LEVEL

Technical Report NAVTRAEQUIPCEN 78-C-0060-3

APPLICATIONS OF ADVANCED EXPERIMENTAL METHODS TO
VISUAL TECHNOLOGY RESEARCH SIMULATOR STUDIES:
SUPPLEMENTAL TECHNIQUES

C. W. Simon
Canyon Research Group, Inc.
741 Lakefield Road, Suite B
Westlake Village, California 91361

DTIC
FEB 25 1981
C

FINAL REPORT   MAY 1978 - MARCH 1980

January 1981

AD A095633

FILE COPY

NAVAL TRAINING EQUIPMENT CENTER
ORLANDO, FLORIDA 32813

81 2 24 036

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NAVTRAEQUIPCEN 78-C-0060-3	2. GOVT ACCESSION NO. AD-A095 633	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Applications of Advanced Experimental Methods to Visual Technology Research Simulator Studies: Supplemental Techniques		5. TYPE OF REPORT & PERIOD COVERED Technical Report 26 May 1978 — 31 March 1980
		6. PERFORMING ORG. REPORT NUMBER CWS-01-80
7. AUTHOR(s) Charles W. Simon		8. CONTRACT OR GRANT NUMBER(s) N61339-78-C-0060 N61339-78-C-0096
9. PERFORMING ORGANIZATION NAME AND ADDRESS Canyon Research Group, Inc. 741 Lakefield Road, Suite B Westlake Village, CA 91361		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NAVTRAEQUIPCEN Project No. 4781-6P5
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Training Equipment Center Orlando, Florida 32813		12. REPORT DATE January 1981
		13. NUMBER OF PAGES 127
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Visual Technology Research Simulator (VTRS)	Research Methods
Advanced experimental methodologies	Human Factors research
Aviation Wide Angle Visual System (AWAVS)	Behavioral research
Experimental design	
Economical multifactor designs	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report is made up of a series of individual papers on techniques to enhance the behavioral research methods being used in the VTRS, or Visual Technology Research Simulator (formerly referred to as AWAVS, or Aviation Wide-Angle Visual System). These methods are applicable to many other topical areas in addition to flight simulation. The techniques discussed, which relate to problems of design, analysis and interpretation, are important addenda to material discussed elsewhere by Simon.   (cont'd)

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE
1 JAN 73
S/N 0102-014-6601 |

In particular, this report supplements NAVTRAEQUIPCEN 77-C-0065-1 (Simon, 1979).

The following techniques are discussed:

a. What to do when the model for the experimental design inadequately represents the empirical data; I. Introduction; II. Lack of Fit Test; III. Transformation; IV. Augmentation.

b. Using Yates' algorithm with screening designs;

c. Analyzing residuals;

d. Identifying the experimental conditions in 2^{k-p} designs when given the defining generators;

e. An economical design for screening interaction effects;

f. Graphic method and internal comparison for multiple response data.

g. The place for replication in economical multifactor research;

h. The significance of tests of statistical significance;

i. Determining the probability of accepting the null hypothesis when in fact it is false;

j. Testing non-additivity in experimental data from a Latin square design.

k. How to include factors with more than two levels in a screening design.

l. Analyzing extra-period change-over designs;

m. Analyzing serially-balanced sequence designs;

n. Design economy when experimental factors selectively affect bi-variate criteria.

Accession For	
NTIS GRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Avail and/or	
Dist Special	

A

PREFACE

The behavioral research conducted under the Naval Training Equipment Center's Visual Technology Research Simulator (VTRS) program consists of experiments to provide the basis for design criteria for flight trainers. Because it is necessary to investigate the effects of a great many simulator features, much attention has been given to the use of experimental methods capable of handling complex multifactor problems. The author of this report, Dr. Charles W. Simon, has devoted the past decade to the study of means to improve the quality and usefulness of behavioral research through the use of methods that are, in many respects, quite different from those typically used by applied behavioral scientists. The critical difference is that with the "new paradigm" (a term used by Simon to refer to the philosophy, strategy and techniques he discusses), variables are examined with a gradually increasing precision as more is learned about their effects. The advantage is that the experimenter is less constrained to investigate only a few variables at a time. He is not forced to hold constant (or allow to vary in some unknown way) other factors that may interact in important ways with those under study.

The experimenter initially looks at many things with the intent of screening out those which are trivial for a particular task. The non-trivial factors are then investigated further until ultimately a sufficiently precise equation is generated and verified. With this approach it would often be possible, with no increase in the amount of data collected, to obtain the same amount of useful information on perhaps 50 variables and their interrelationships as would ordinarily be obtained on five variables using traditional methods. Predictions from the laboratory to the field can then be made with greater confidence, and a quantitative data base is established which can be augmented easily.

A report recently published as NAVTRAEQUIPCEN 77-C-0065-1 (Simon 1979) summarized the ways in which the methods he advocates should be applied to the VTRS program. The present report supplements that document by providing additional information to aid in the design and interpretation of multifactor experiments. The information presented here was required in preparation for a screening experiment recently completed on aircraft carrier landing performance. A description of this research will appear as NAVTRAEQUIPCEN 78-C-0060-7. The practical implementation of the techniques will be discussed in that report.

There are many *fields of experimental psychology* for which the approach advocated by Simon can result in more useful information obtained at a lower cost. This report will therefore be applicable to a wide range of research topics besides flight training. Because it assumes that the reader is familiar with Simon's earlier work, *it should be regarded as a* "companion piece" to NAVTRAEQUIPCEN 77-C-0065-1, which provides much of the background information a new reader would require.

STANLEY C. COLLYER
Scientific Officer

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION I

INTRODUCTION

The Naval Training Equipment Center has built a Visual Technology Research Simulator (VTRS, formerly referred to as AWAVS) composed of a cockpit, a wide-angle visual system, and a six-degrees-of-motion system. These combine into a versatile device for studying the effects of equipment parameters in the context of pilot training. The large number of parameters that must be investigated requires the use of experimental methods that permit studying many factors economically. A discussion of the philosophy, strategy, and techniques being employed in much of the research conducted on this program has been provided elsewhere (Simon 1979).

This report is made up of a series of individual papers on different techniques needed to enhance the methodologies that are to be used in the VTRS human performance experiments. In the series of reports by Simon (1970 - 1979) on a new paradigm for psychological research, the holistic approach to systematic experimentation is proposed and the strategies and techniques for accomplishing this are described. While the basic tools required to employ the "new paradigm" (Simon, 1977b) in the VTRS program are available, there are still techniques that need to be understood in detail to supplement the use of those described in the original documents.

As a part of this year's effort, supplemental procedures for the design, analysis, and interpretation of economical multifactor experiments were sought. The relevant ones are described here. None is original with this investigator. They have been included here to reduce the time required to search them out, to read and collate related source material, and to relate them to the "new paradigm." After using this report to obtain a basic understanding of these techniques, the reader is encouraged to read the original material.

The following techniques are discussed:

a. WHAT TO DO WHEN THE MODEL FOR THE EXPERIMENTAL
   DESIGN INADEQUATELY REPRESENTS THE EMPIRICAL DATA

       i   THE PROBLEM
     ii   LACK OF FIT TEST
   iii   TRANSFORMATION
    iv   AUGMENTATION

b. USING YATES' ALGORITHM WITH SCREENING DESIGNS

c. ANALYZING RESIDUALS

d. IDENTIFYING THE EXPERIMENTAL CONDITIONS IN 2^{k-p} DESIGNS WHEN GIVEN THE DEFINING GENERATORS

e. AN ECONOMICAL DESIGN FOR SCREENING INTERACTION EFFECTS

f. GRAPHIC METHOD AND INTERNAL COMPARISONS FOR MULTIPLE RESPONSE DATA

g. THE PLACE FOR REPLICATION IN ECONOMICAL MULTIFACTOR RESEARCH

h. THE SIGNIFICANCE OF TESTS OF STATISTICAL SIGNIFICANCE

i. DETERMINING THE PROBABILITY OF ACCEPTING THE NULL HYPOTHESIS WHEN IN FACT IT IS FALSE

j. TESTING NON-ADDITIVITY IN EXPERIMENTAL DATA FROM A LATIN SQUARE DESIGN

k. HOW TO INCLUDE FACTORS WITH MORE THAN TWO LEVELS IN A SCREENING DESIGN

l. ANALYZING EXTRA-PERIOD CHANGE-OVER DESIGNS

m. ANALYZING SERIALLY-BALANCED SEQUENCE DESIGNS

n. DESIGN ECONOMY WHEN EXPERIMENTAL FACTORS SELECTIVELY AFFECT BI-VARIATE CRITERIA

RELATING THE CONTENT OF THIS REPORT TO PREVIOUS REPORTS

The new paradigm for research on equipment design developed by Simon (1970-1979) emphasizes the importance of a multifactor approach involving "all" critical parameters of a particular task and provides the practical and economical strategies and techniques for accomplishing this. Much of the material written about the basic approach has been presented as if it were a constant, unvarying process. In practice, however, the comprehensiveness of the approach and the vagaries of humans performing complex tasks in less than optimum environments makes any "cookbook" approach inadequate. The investigators must be prepared to handle variations upon the basic approach and to deal with the complexities of the problem as it exists in the real world. If they cannot, the holistic approach for behavioral research will not be successful. The material in this report is intended to supplement the material already written in order to better prepare the investigator for conducting experiments on VTRS and similar programs. The discussion below relates the sections in this report to previous materials dealing with problems of design, analysis, and interpretation.

Design

The term "design," will be used here in the limited sense, to refer to the selection of coordinates in a multifactor space where performance data is to be collected. In the basic approach to economical multifactor experiments, data is collected in blocks beginning with the points which form designs of lower order so that a sequential build-up of designs of increasing order will be possible. By testing after each block of data is collected, the investigator can develop a polynomial of appropriate order to be fitted "as close as possible" to the true unknown response function while using as few experimental runs as is consistent with other objectives. The basic design used in this approach is a 2^{k-p} fractional factorial, which is modified to achieve the experimental goal. Items a-iv, d, e, g, k, and n are concerned with some of the most frequently encountered modifications.

After the design has been expanded to approximate a second order model, it is uneconomical to collect -- if needed -- all of the points required for a complete design of higher order. Instead, the investigator must know how to augment the data collection space with points that will isolate specifically chosen sources of variance (Item a-iv). Sources that supply unusual experimental data collection plans may provide only the "defining generators," that is, a succinct coded description of the design. An investigator must know how to determine which experimental conditions (i.e., the coordinates of the experimental space) make up this design (Item d). Ordinarily, investigators are interested in identifying which factors are the most important and most designs -- particularly the economical multifactor designs -- are constructed to reflect this interest. When interactions are expected to be predominant and a large number of factors are being investigated, the investigator who is familiar with an economical plan for screening interaction effects can save much time and effort (Item e). Most psychologists replicate basic experimental designs almost automatically, whether or not it serves a useful purpose and without regard for the extra data collection costs involved. When large multifactor experiments are to be performed with reasonable economy, the investigator must understand when replicating is and is not necessary and what more economical alternatives to complete replication are available (Item g). While the two-level design is usually suitable for most screening studies at the beginning of a large multifactor investigation, there are times when the investigator may wish to examine one or two factors at three (or even four) levels when the first block or two of data is being collected. He must be aware of when this is reasonable, what alternatives are open to him, and if he decides to go ahead, how to fit the three-level factor economically into the two-level basic design (Item k). For truly holistic experimentation, multifactor experiments will commonly require

multivariate criteria. The investigator should be aware of those special circumstances when the experimental data collection can be made more economical (Item n).

Analysis

Too often, the analysis of behavioral research data has been limited to a routine, computerized treatment. This is not sufficient when the data is as rich in information as that which comes from a multifactor design. Furthermore, when "economical" designs are used, it is necessary not only to analyze for content, but also for data quality in order to avoid potential misinterpretations. Items a-iii, b, c, and f deal with topics related to the analysis of such data.

In some cases, proper data analysis may be substituted for the additional data collection required by a more complex design. Before one adds a new block of data to isolate higher order interaction effects as prescribed by sequential design strategy described earlier, it is more economical to examine transformations of the original performance data to see if a simpler model can be effected without requiring more data to be collected. With multifactor designs, the investigator must be knowledgeable of special techniques necessary to optimize transformations across all variables (Item a-iii). When 2^{k-p} designs are used as blocks in the sequential strategy, using Yates' algorithm for data analysis can prove to be more effective than computerized regression routines when the number of factors are more than the computer system can handle. With complete 2^k factorials, the Yates' algorithm is easy to use; with fractional 2^{k-p} designs, including the special case of the robust screening design, some translation of the results is required. The investigator must be able to make this translation in order to interpret his data (Item b). Analyzing the residuals between obtained and estimated performance scores provides valuable information for the investigator who wishes to understand the quality of his data, to decide what the next step in the experiment should be, and to interpret the existing data. Most psychologists are not aware of the usefulness of this type of analysis and should be (Item c). Holistic research is expected to deal with a complex world -- multiple independent variables and multiple criteria. Graphic methods, so helpful in understanding data from single-criteria experiments, can also be useful when multiple criteria are employed. This latter analysis however, is more complicated to perform and to interpret; the investigator needs simplified explanations of both (Item f).

Psychologists have employed Latin square designs for many years to isolate the effects of treatments, subjects, and trials when the same subject is tested across all treatments. At the same time, almost none have evaluated their data to see if it meets the assumptions required to use the design (Item j), nor have used the analysis to make

full use of this data collection plan to isolate effects that might be carried over from one treatment to the next, a linear transfer effect (Items 1 and m). If one can justify examining only linear transfer effects, this class of design would be an economical means of studying transfer of training using a small group of subjects or only a single individual. Information on how to analyze these two types of designs is not readily available nor always clear, and so it is provided here (Items 1 and m respectively).

Interpretation

In addition to design and analysis, the third leg of any research effort is the interpretation of the data. This differs from the analysis since the former merely organizes and summarizes the data while the latter considers the practical implications of the results. Some aids to inter-pretation can be found in the sections on design and analysis.

For the paradigm for large-scale multifactor research, tests of lack of fit are one means of deciding whether or not enough data has been collected to write an equation that adequately approximates the experimental space. This lack-of-fit test should be made after each block of data of higher order has been collected (Item a-ii). Because of the fanatic and sometimes frenetic reliance that psychologists place on "tests of statistical significance" in the interpretation and evaluation of their experimental data, it is important that the ways this test has been misused and misinterpreted be understood by the practicing experimenter. A large number of papers have appeared in the behavioral science literature spotlighting these deficiencies, without, it seems, having done much to reduce them. A summary of the facts and fallacies that surround this procedure (Item h) should sensitize the psychologist who uses this test to its limitations as well as to the way it has frequently been misused.

The techniques described in this report are of limited value in isolation. On the other hand, they are important addenda to the material already discussed by Simon (1970-1979) elsewhere and have specific applications for a properly conducted VTRS program as well as similar research projects in the future.

SECTION II

WHAT TO DO WHEN THE MODEL FOR THE EXPERIMENTAL DESIGN INADEQUATELY REPRESENTS THE EMPIRICAL DATA

THE PROBLEM

One desirable feature of a "good" experimental design for mapping the response surface of an experimental region is that it includes points which form designs of lower order so that a sequential build-up of designs of lower order is possible (Box, 1964). This feature greatly enhances the economy of data collection since the information collected early in the sequence can be used to identify critical factors, permitting factors contributing only trivial effects to be dropped before the response-fitting phase begins. In general, it will be true that relatively few factors will be needed to account for most of the performance variability in a specific task. Quite frequently, a Resolution IV design will provide nearly all of the information required for factor identification and a second-order design will adequately describe most response surfaces of human performance. But what happens when either of these statements are not true? What procedures must an investigator employ then?

In the discussion that follows, alternative actions available to an investigator when these situations occur are described. Since the solutions for following up on screening designs and for expanding the *response surface* involve the same or similar considerations, these two problem areas are treated together here.

Alternative Actions

Had the investigator anticipated the possibility that higher-order effects might be present, he might have <u>started with a particular experimental design</u> capable of expanding to the desired order. Thus there are Resolution III, IV, and V fractional factorials capable of being built from lower order designs. Similarly there are second and third-order response surface designs that can be built sequentially from lower order designs. As a general principle, however, Resolution V fractional factorials and sequential third-order response surface designs tend to be uneconomical, requiring more data collection than is probably necessary. For this reason, it is less likely that the investigator -- even if he anticipates the need -- will prefer to employ this alternative.

When a test of lack of fit reveals the presence of higher-order effects not yet included in the polynomial, the investigator should first <u>inspect</u> his data for deviant values from irrelevant sources. Unusual values not associated with the true intent of the experiment may occur if a subject fails to respond in accordance with instructions,

if equipment fails or otherwise "misbehaves," and if data is recorded (or analyzed) incorrectly. These and similar disturbances in the data can create results that mathematically appear as higher-order interaction effects. Any such outliers therefore must be detected, not only to reduce the chances of drawing erroneous conclusions but also to prevent the investigator from being misled into collecting more data unnecessarily to isolate these quasi-interactions.

If there is no reason to suspect the accuracy of the data, the investigator may consider transforming the data to reduce or eliminate higher-order effects before considering the less economical approach of collecting more data. Psychologists have more frequently employed data transformations to meet the assumptions of statistical analysis than they have to simplify the regression model. If a simpler model can be effected with transformations so that higher-order effects are for all practical purposes eliminated, then the amount of data needed to approximate the correct model is reduced. The difficulty in applying this approach is in selecting the proper transformation or transformations. Then, too, some types of non-additivity can never be eliminated with transformations.*

When transformations fail to simplify the data, the investigator has no other choice but to collect additional data to isolate the effects that account for the lack of fit. Exactly what is best to do in this case is not always obvious when economy must be a primary consideration. It takes a considerable amount of additional data to change a Resolution IV screening design to Resolution V. No standard methods are available to expand a second-order central-composite design when the need for a third-order model is indicated. However, there are procedures for selecting a limited number of data points that will isolate only the effects of greater interest.

Major topics to be discussed in the sections that follow include:

- Lack of fit tests

- Multivariate transformations

- Augmentation techniques

---

* An ounce of prevention is worth a pound of cure. Selecting the scaling should be a major effort of the pre-experimental planning phase.

TESTING THE FIT OF THE MODEL

In the 2^{k-p} screening design, lack of fit refers to the need to include interactions in what otherwise would be a first degree polynomial. Their presence may be inferred when a string of interactions in the Resolution IV designs shows a non-trivial effect.* There is no need to test for curvature during the screening phase since the factors are only at two levels. Since these designs are not replicated there are no degrees of freedom for extimating error. However, a test of the statistical significance of potentially interesting effects can be approximated using order normalized plots and estimated error variances (Daniel, 1976; Simon, 1977a).

With quantitative factors, the goal ultimately will be to estimate a response surface. Lack of fit tests are needed to see if the completed screening data, a first degree polynomial plus critical interaction terms, adequately fits a linear surface or whether the curvature of the surface must also be accounted for. To determine whether quadratic terms are needed to approximate the response surface, it is necessary to add center points to the screening design. If it can be anticipated that this step will eventually be taken, it is better to do so when the screening data is being collected, rather than later (Simon, 1977a).

To determine whether or not quadratic terms are needed to fit the response surface, the investigator would compare the average performance of all points in the hypercube against the average performance at the center, i.e.,

$$(\bar{x}_{hc} - \bar{x}_c).$$

This measure of overall curvature is equal to the sum of the estimated coefficients of the quadratic terms. Whether or not it is larger than would be expected by chance is determined by the magnitude of its ratio to the estimated error.

To determine whether a second degree polynomial is adequate, tests can be made in the conventional manner. The sum of squares for a lack of fit term (with one degree of freedom) can be obtained by subtracting from the total sum of squares all of the sums of squares for linear, quadratic, and interaction terms, along with the sums of squares for

---

* The investigator must also be alert to the possibility that within a string two large effects with opposite signs could yield a trivial sum.

center points (error) and blocks -- if present. An F ratio
of overall lack of fit variance to error variance can be
obtained. Draper and Herzberg (1971) describe a way of
partitioning this overall estimate into lack of fit due to
interaction and that due to cubic effects (see Simon, 1977a,
pp 169-171).

Box, Hunter, and Hunter (1978, pp 522-523) show a short-
cut method of testing the adequacy of the second-order model.
They point out that "if the surface is exactly quadratic in
this direction [of a single dimension], it can be shown that
the estimate of slope obtained from the axial points [of
that dimension] will be the same as that obtained from the
factorial points [of that dimension]." The slope, $\underline{m}$, of a
line is equal to

$$m = \frac{y_1 - y_2}{x_1 - x_2} .$$

Let us use this to test whether the second-order central-
composite design adequately fits the data or if there is a
third-order component.

To obtain the slope of the axial (star) points (m_a), we
substitute:

y_1 = the performance at axial point $+\alpha_{x_i}$

y_2 = the performance at axial (star) point $-\alpha_{x_i}$

x_1 = the coded value of $+\alpha_{x_i}$

x_2 = the coded value of $-\alpha_{x_i}$

To obtain the slope for the factorial (cube) points (m_f), we
substitute

y_1 = the average performance at factorial points $+1_{x_i}$

y_2 = the average performance at factorial points -1_{x_i}

and

x_1 = the coded value +1

x_2 = the coded value -1

The difference between the two slopes,

$$(m_a - m_f)$$

supplies an estimate of the sum of the third-order coeffi-cients, i.e.,

$$\Sigma \beta_{iy2}$$

where y may equal both i and non-i and i equals each of 1 to K factors. In the third-order polynomial these are the combined coefficients for third degree terms,

$$x_i x_y^2 \quad \text{and} \quad x_i^3.$$

TRANSFORMATIONS TO REDUCE LACK OF FIT

When the derived polynomial fails to fit the data and there are not enough degrees of freedom remaining to expand the model (as is the case with classic second-order central composite designs and 2^{R-p} screening designs), the investigator should first try to simplify the relationship by transforming the data. This is not a completely foreign procedure to psychologists who have used logarithmic transformation to linearized the relationship between psychological and physical scales. If the response surface of the transformed data can be approximated by a lower-order polynomial than had been possible with the untransformed data, then there will be no loss of information and the cost of collecting additional data will have been avoided. The investigator faces the task of deciding what is the best transformation to use.

Selecting a transformation is not a simple task. Too often the process has been oversimplified and treated in cookbook fashion. Sometimes simple generalizations are made verbally, such as: "A reciprocal transformation should be used when reaction-time scores are involved." In other cases, the arithmetic relationship between the mean and the variance of the data is used to select the transformation, e.g., "When the variance (s^2) equals the mean multiplied by some constant, (kM), a square root transformation should be used". These are monotonic transformations; that is, they handle relationships with "one-bend" in them. Few psychologists ever consider "two-bend" transformations, the one exception being the arcsin transformation for handling percentage data. The most limiting feature of this type of treatment of transformations is that they deal only with the one-independent, one-dependent variable case.

The selection of transformations becomes more complicated when there are multiple independent variables. If their relationships with the dependent variables differ, we will not be able to transform the response data but must find the appropriate transformation for each independent variable. This, however, will destroy the orthogonality among the independent variables. Additional problems can arise if the transformation destroys the normality of the distribution and/or the homogeneity of the variance. Luckily, many transformations that correct one deficiency in the data correct another; still it is important to realize that this is not always the case.

More sophisticated treatments of data transformation have noted that what are often viewed as independent methods of modifying the data, e.g., reciprocal, logarithmic, square root, square, and others used with monotonic data, are actually members of a common family that can be represented by a single equation, which varies the "strength" of the

transformation as parameters of the equation are changed.

Box and Cox (1964) present one such general transformation equation

$$w = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ \log y & (\lambda = 0)* \end{cases}$$

where y is the response data and lambda (λ) is the parameter to be varied. As lambda takes on the values from -1 to +1 (and the range can be greater), the equation becomes equivalent to a number of transformations commonly employed by psychologists (see Table 1).

TABLE 1.   RELATION AMONG COMMON MONOTONIC
TRANSFORMATIONS, SIGMA-MEAN RATIOS, AND LAMBDA

Transformation	Relation	Lambda	Application
Reciprocal	$s = kM^2$	-1.0	Reaction time
Reciprocal square root	$s = kM^{3/2}$	-0.5	-
Logarithm	$s = kM$	0 *	Positively skewed data; sigma
Square root	$s^2 = kM$	+0.5	Frequencies
No transformation		+1.0	

*This is a convention that allows a plot of results against lambda to be continuous.

A Practical, Empirical Approach to Multivariate Transformation

Draper and Hunter (1969) propose a rather straight-forward data-inspection technique for selecting the desired transformation of the dependent or independent (or both) variables. Their procedure takes advantage of a high-speed computer to analyze the data after different transformations have been effected. The desired transformation would be selected after a visual inspection of the results from these several analyses are properly plotted.

To systematize this process, Draper and Hunter made use of the equation for the family of monotonic transformations developed by Box and Cox. A computer is programmed to iteratively change the value of lambda in the equation in regular increments and to perform an ordinary analysis of variance after each transformation.* The results are plotted on a graph with lambda values along the abscissa and the results of the ANOVA along the ordinate. All sources of variance are plotted on the same graph. Since lambda can take on any value, the plots at the different lambdas can be connected into a continuous function. Given these functions the investigator can select the value of lambda (and thus the transformation) that best meets his requirements.

To illustrate this procedure, several examples taken from Draper and Hunter's paper will be described briefly. The reader is encouraged to refer to the original paper for additional details.

In their first example, they show how a transformation of the dependent variable can be selected to eliminate an interaction. From the data collected in a two-factor, 3 x 4 factorial design, mean squares and F-ratios are determined for main effects A and B and their interaction AB. The error variance is also obtained. Using the Box and Cox equation, Draper and Hunter transform and analyze their data using all conditions of lambda between -2 and +1.2 in steps of .01. Next they plot against lambda the magnitudes of the F-ratios** for the three effects. They also plot a measure of "inhomogeneity" to see if the evenness of the within-call variances is disturbed severely by the transformations (see their Appendix, pp 38-39, for the equations needed to

---

* There is no reason why other transformations might not be employed.

** The F-ratios should be used whenever the dependent variable is transformed, since each transformation of y yields a different total sum of squares. This is not the case when only the independent variables are transformed; then a plot of the mean squares is appropriate.

calculate this measure). The complete plot is shown in Figure 1. Inspection of this figure reveals the lambda (and thus the transformation) that minimizes the interaction, maximizes factor effects, and keeps the inhomogeneity within acceptable bounds.

In this example, Draper and Hunter decided to consider inhomogeneity values lying within the 95 percent one-tail confidence limits. This places the candidate lambda between -1.75 and -0.53. The two main effects are maximized at -1.35 and -1.25, respectively, and the interaction is minimized at approximately -0.60. They recommend a lambda of -1.00 as a sensible compromise. Inhomogeneity at this point is not a minimum, but it is not excessively high. A lambda of -1.00 in the Box and Cox equation means that a reciprocal transformation should be used.

In their second example, Draper and Hunter show how they selected a transformation to linearize the results of a 3^3 factorial experiment analyzed by "standard response surface methods." In this case, the total variance is partitioned into linear, quadratic, and residual variances. To find the transformation of the dependent measure that would enable the three dimensional surface to be represented by a first-order equation, they plot the F ratios for the linear and quadratic terms against values of lambda. They suggest that since the goal is to maximize the linear component and minimize the quadratic, one might also plot the ratio of the linear mean square over the quadratic mean square against lambda. Inspection of the graphic plot (Figure 2) shows that when this ratio is maximum, the lambda is essentially zero which means the desired transformation is logarithmic.

In a third example, Draper and Hunter illustrate how this graphic method can be used to select a transformation if both independent and dependent variables are transformed. They state as a general principle that when multiple variables are to be transformed, it is not appropriate to transform the variables one at a time, an often used method; instead all must be transformed simultaneously (see Hill, 1966). The particular data used for this example was from an undesigned experiment. The problem was to fit the data with the simplest form of the regression model of the form

$$E(w) = \beta_\alpha X^\alpha$$

where w is the Box-Cox transformation with λ the unknown. In this case the equation was modified; $(y+1)^\lambda$ was substituted for y^λ. Since both independent (y) and dependent (x) variables must be transformed, we must determine the values of two unknown transformations, α and λ. The sum of squares for this model can be partitioned into that due to the regression, the lack of fit, and pure error. We need to find
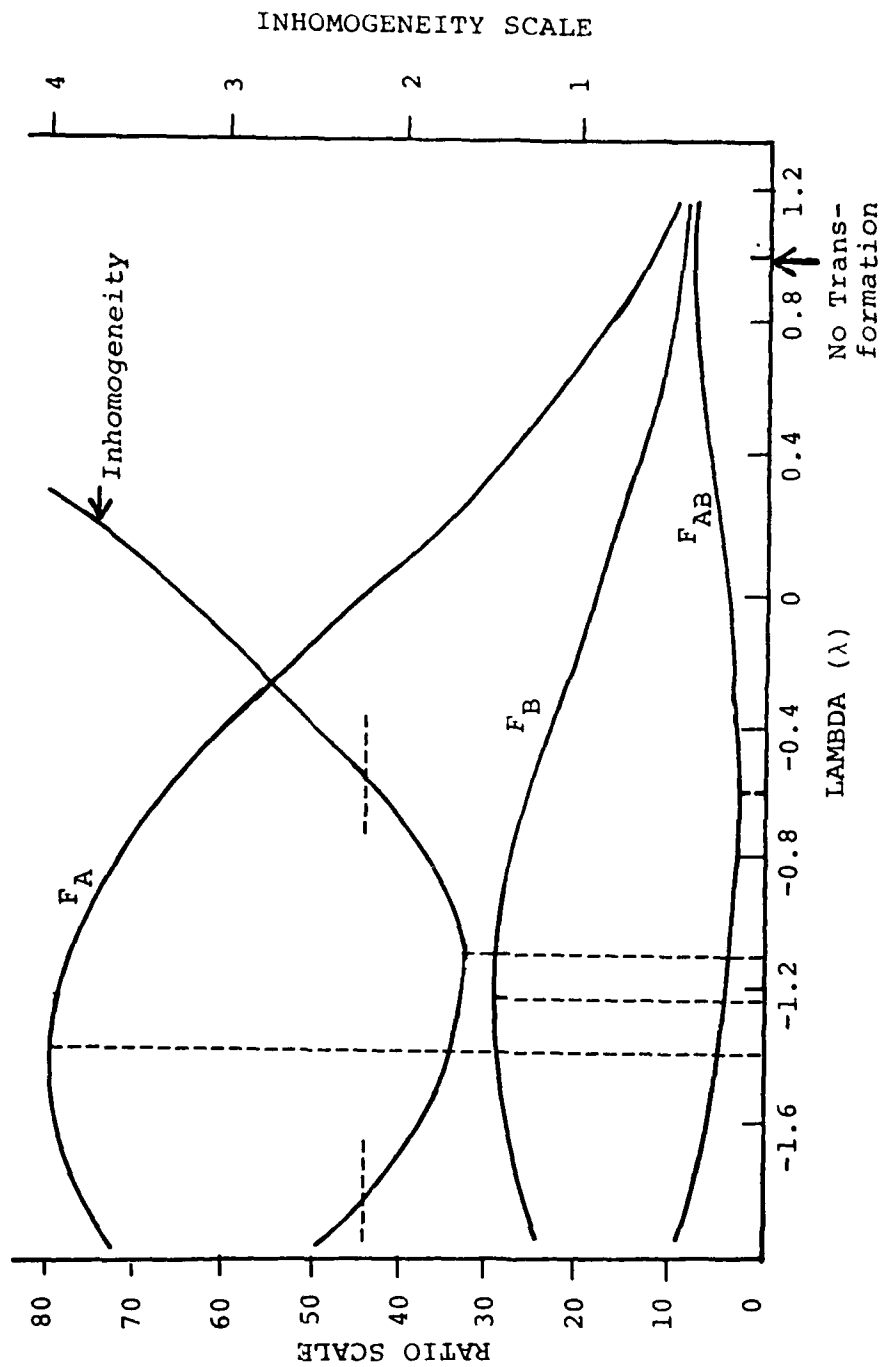
Figure 1. Plot to Select Transformation to Get Rid of an Interaction
[From Draper and Hunter (1969)]
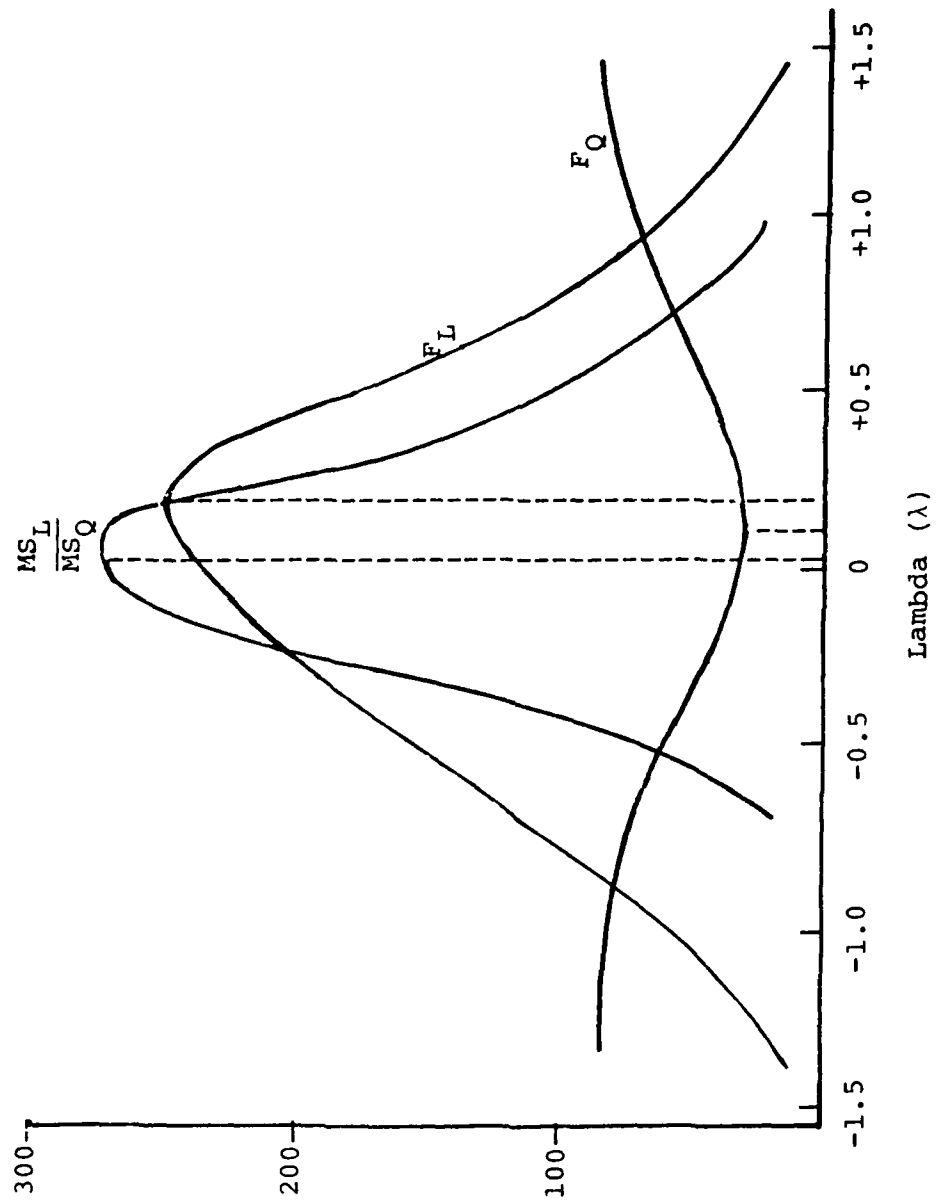
Figure 2. Plot to Select Transformation to Linearize a Response Surface
[From Draper and Hunter (1969)]

the minimum ratio of the mean square for the lack of fit over the error variance, that is, the minimum F-ratio for lack of fit.

Figure 3 shows a contour plot of the F-ratio (ℓ) for lack of fit plotted on a two-dimensional graph of alpha versus lambda. Draper and Hunter also calculate the measure of inhomogeneity for the lambda dimension; the vertical lines show the acceptable region for this measure. The preferred transformation is where the lack of fit contour is at a minimum. This is located where lambda equals one (signifying no transformation) and alpha equals 1.5. Thus the simpler model is

$$E(y) = \beta x^{\frac{3}{2}}$$

to provide an additive representation of this set of data.

Draper and Hunter make several general suggestions regarding the use of this graphic plot technique for selecting transformations. First, since there are usually a band of choices that can be made, better decisions can often be made given several sets of data collected on the same problem. Second, while the empirically-derived transformations have a pragmatic value and can be found, they must not totally divert attention from eventually developing a theoretically-based model. [Simon says: That may be easier to do in the physical than in the behavioral sciences.] Third, and most important, they recommend that once a transformation has been selected and the data analyzed, it be subjected to a residual analysis (see Daniel, 1976; Draper and Smith, 1968; Box, Hunter, and Hunter, 1978). This residual analysis is first performed on the transformed data using the higher-order polynomial model. Then if it appears that the simpler model is appropriate, the transformed data is reanalyzed with the higher-order terms omitted and the residual analysis is repeated on these results. Finally, they suggest that a degree of freedom should be removed from the residual for each parameter that is transformed.

It should be noted in closing that transformation will not simplify all forms of data. When interactions are disordinal (or intrinsic), they cannot be eliminated by transforming the data. Transformations will only simplify ordinal interactions and curvilinear effects. Luckily, these occur the most frequently in human performance data, and so transformations do much to reduce the need to collect additional data when lack of fit occurs.

When interactions are eliminated by means of data transformation, care must be taken not to misinterpret the results. If a critical interaction effect, found in the original analysis, is no longer found after the data has been trans-
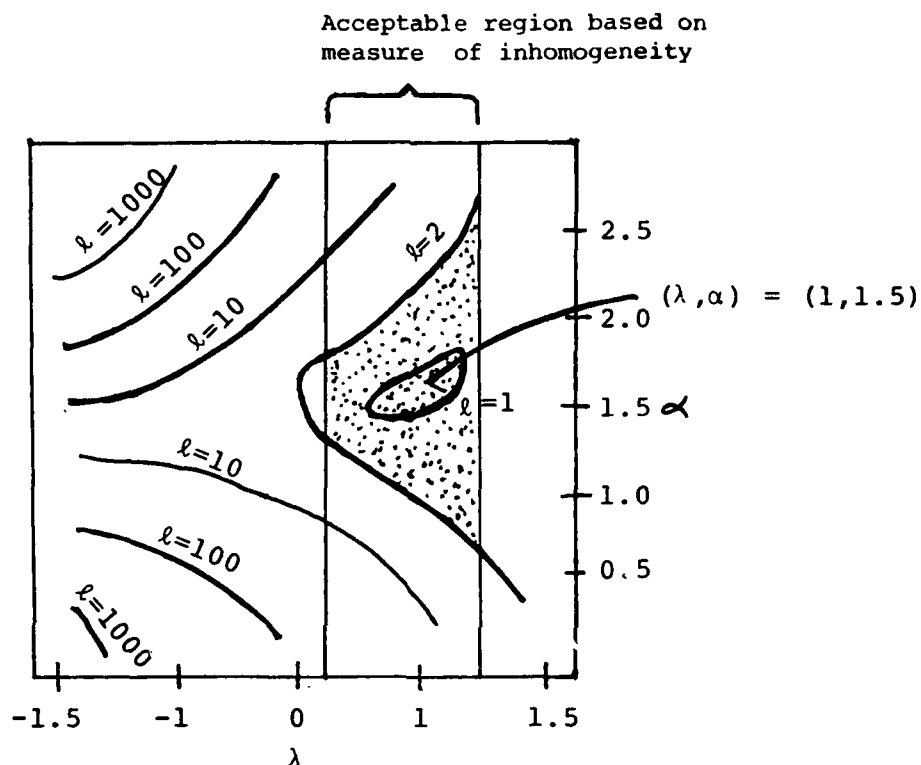
Figure 3. Plot to Select Transformations when Changing Both
X and Y*
[From Draper & Hunter (1969)]

---

*α is transformation for X, and λ is transformation for Y.

formed, there is no contradiction in the true interpretation
of data. It is not enough to say, as too many psychologists
do, that there is or is not a statistically significant
interaction without relating it to the measurement scale that
is involved. Results of psychological experiments are
"situation specific", and this includes the characteristics
of the data and its analysis as well as those of the subjects,
equipment, environment, and time dimension.

AUGMENTATION

When transformations fail to simplify the data to the
level at which it can be approximated adequately by a lower-
order (i.e., 1st or 2nd) model, additional data must be
collected. There are a number of occasions as the sequen-
tial development of the "new paradigm" unfolds when the
investigator will be faced with the problem of deciding what
augmentation data collection plan he must use.

There are two occasions on which the augmentation design
is fairly stereotyped and can be derived rather mechanically.
These are:

1.   When the second block must be added to the
      original Resolution III screening design in
      order to isolate main from two-factor
      interaction effects (see Box and Hunter, 1961;
      also Simon, 1973, pp 105-116).

2.   When axial points must be added to the hyper-
      cube of a central-composite design in order to
      approximate a second-order response surface
      (see Box and Hunter, 1961; also Simon, 1970;
      1973, pp 131-139).

There are, however, other occasions when more data must be
collected but the rules for selecting the data points are
less well-defined. The investigator in these cases must
exercise his judgment and skill. The more important of
these occur in three segments of the research program.
Additional data may have to be collected:

1.   To isolate critical two- or three-factor disordinal
      interactions from strings obtained from a
      Resolution IV design during the screening phase.
      This augmentation is necessary to avoid the
      risk of ignoring influential factors, the
      effects of which might be found in their
      interaction with one or more other factors.

2.   To isolate all critical two-factor inter-
      actions aliased in strings of a Resolution IV
      design when screening data is to be used to
      complete the "hypercube" portion of a central-
      composite design. If all critical two-factor
      interactions are isolated from main effects and
      one another, then the plan is equivalent to a
      complete Resolution V design in which all two-
      factor interactions are isolated from main effects
      and one another. This equivalence is achieved
      with a reduced data collection effort. This
      reduced design will be referred to as a
      Resolution V- design.

3. To isolate third-order terms when the second-order model does not adequately fit the response surface. This is necessary to minimize the prediction bias based on an inadequately developed polynomial.

So for whatever reason one isolates the interactions in strings, more data most be collected. The goal is to do this as inexpensively as possible. Two basic approaches are employed: 1) To isolate all interactions within the string, or 2) To attempt to "guess" which interactions are critical and probe with a few data points to verify the hypothesis. Which approach will be used depends on: 1) A priori information regarding potentially important interactions; 2) The length of the aliased strings; 3) The cost and time restrictions on further data collection; 4) The precision with which the effects must be estimated; and 5) The damage that could occur if an effect is neglected.

Isolating All Interactions within a String

Techniques for isolating all the interactions within a string have been described by Box, Hunter, and Hunter (1961) and Daniel (1976). The example that follows was taken from Box, Hunter, and Hunter's book (1978, pp 414-416). First the method of selecting the new experimental conditions is described; then an example is given to show how the new data is combined with the data from the original block.

We shall assume that data has already been collected to complete a 2_{IV}^{8-4} screening design in which there are eight candidate factors and 16 experimental conditions. The design was actually composed of a Resolution III design plus its fold-over. The values of the mean, the eight main effects (aliased with three-factor interactions), and seven strings of four two-factor interactions are all given in Table 2. Among the effects, Factors C, E, and H appear critical along with the two-factor interaction string, (AE). This string represents the combined effect of the four aliased inter-actions: (AE + BF + CH + DG).

To select the conditions required to isolate the inter-actions, N additional conditions are required to form an orthogonal block for isolating N interactions in the string. In our example, N = 4. A pattern of signs in an inter-actions-by-conditions matrix must be determined that satisfy the orthogonality requirement. One such pattern is shown on the left side of Table 3. If the interactions have these particular signs for each condition, then the main effects must have signs such as those shown on the right side of the same table.

30

TABLE 2.   CONTRIVED RESULTS FROM BLOCK I DATA

A	B	C	D	E	F	G	H
-.7	-.1	5.5	-.3	-3.8	-.1	.6	1.2

(AB)	(AC)	(AD)	(AE)*	(AF)	(AG)	(AH)	Mean
-.6	.9	-.4	4.6	-.3	-.2	-.6	19.75

[*(AE) = AE + BF + CH + DG]

TABLE 3.   SELECTING THE CONDITIONS TO
ISOLATE INTERACTIONS IN STRINGS

Interactions to be isolated					Block of conditions capable of isolating the interactions to left								
	AE	BF	DG	CH		A	B	C	D	E	F	G	H
1)	+	-	-	+	1)	-	+	+	+	-	-	-	+
2)	+	+	-	-	2)	-	+	-	-	-	+	+	+
3)	+	-	+	-	3)	+	+	-	-	+	-	-	+
4)	+	+	+	+	4)	+	+	+	+	+	+	+	+

TABLE 4.   SIGN PATTERN AND RESULTS FOR CRITICAL
EFFECTS OF THE AUGMENTATION BLOCK

	...Critical Effects....								....Results...		
	M	C	E	H	AE	BF	DG	CH	Y	Y'	Y"
Coefficients from Block I		2.75	-1.9	.6	[½ string effect = 2.3]						
Run 17	+	+	-	+	+	-	-	+	29.4	24.15	3.2
Run 18	+	-	-	+	+	+	-	-	19.7	19.95	-1.0
Run 19	+	-	+	+	+	-	+	-	13.6	17.65	-3.3
Run 20	+	+	+	+	+	+	+	+	24.7	23.25	2.3

Neither set of signs are unique, but all must satisfy the relationships found between the main and interaction effects. Thus, if we make the sign of Factor A a minus (actually a -1), then in order to have the sign of AE interaction a plus (actually +1), it is necessary that E also be assigned a minus. Similarly, if BF is to have the sign of minus in the first condition, then either B must be plus and F, minus, or vice versa. In this example, the former pattern was used. Once the signs have been determined for all of the main effects, the experimental condition can be identified by letters with the plus signs. For the first condition in the example in Table 3, therefore, with B, C, D, and H showing the plus sign, the experimental condition will be designated bcdh, indicating which factors are to be set at the high (or +1) level for that condition.

To combine old and new data, let us construct Table 4. We will begin by calculating the coefficients for the Block I (original) data. These are equal to one-half of the value of each effect. Thus the effect of Factor C equals 5.5; its coefficient equals 2.75. Since the effects have not yet been isolated, we cannot know the coefficients for the individual interactions in the string; only the coefficient for the combined string can be determined. In this example, it would be one-half of 4.6, or 2.3. These values along with the appropriate sign pattern are laid out for the four additional runs, numbers 17 through 20, as shown in Table 4. To the right of this table are three columns, designated Y, Y', and Y".

Column Y contains the performance score obtained for each of the four new runs. (Incidentally, note that for Run 20 the signs of the interactions in the string are all plus, corresponding to the levels in the original string.)

Column Y' contains the values after the Y values have been corrected for the known main effects. Using Run 17 to illustrate this step, we substitute coefficients and terms in the following equation:

$$\text{Mean} + \beta_C C + \beta_E E + \beta_H H + \beta_{(AE)} 0.5(AE) = \hat{Y}$$

thus

$$\text{Mean} + (2.75)(+1) + (-1.9)(-1) + (.6)(+1) + [0.5(AE)](+1) = 29.4$$

and consolidating these values, we obtain:

$$\text{Mean} + [0.5(AE)] = 29.4 - 5.25$$

or Y'_{17} equals:

$$\text{Mean} + [0.5(AE)] = 24.15$$

The technique described above can be used to isolate three-factor interactions in strings in the same way it was used to isolate the two-factor interaction effects in strings.

Augmenting Designs with Incomplete Blocks

The investigator may, for various reasons, not wish to add a complete block of new data. In a screening design or the factorial portion of the central-composite design, he may want to use only a few additional data points to probe and crudely test his hypothesis that a particular interaction within a string is accounting for most of the observed variance within the string. Too much "guessing" of this sort can turn out to be more expensive than a more formal approach, but it must be considered as a viable alternative when there are a great many effects within strings and when the clues are available and strong (Simon, 1973, pp 112-115).

In the case of a response surface, an investigator may wish to add some data points where the slope of the surface is steepest, and/or changing rapidly, to improve the precision of the estimates. Or he may wish to add star points in one corner of the original fractional factorial plan to develop a non-central composite design.

In both cases, adding only a few points will often destroy the orthogonality of the design, a condition that is more important in the identification phase than in the response surface phase of the research. In the latter case, we have little interest in individual factors or their effects and are concerned primarily with the overall shape of the multifunction surface. A regression analysis can be used to incorporate the new data into the earlier data. If it is important, a few additional points might be employed to improve the orthogonality using the technique proposed by Dykstra (1971; also see Simon, 1975, pp 26-30).

When new conditions are tested at periods of time far removed from that when the original data was collected, one must be careful of shifts in the performance level due to external irrelevant sources. Orthogonal blocking techniques would ordinarily be employed to handle these situations. Where complete blocks are not involved, however, we might follow Hebble and Mitchell's (1972, p 771) suggestion of using a blocking term in the regression equation which already has a constant, β_o, in the model. A "dummy variable" is created by assigning a zero value to each condition in the original design and a value of one to the new conditions. Regression techniques would be used to analyze the combined data. When this is done, the model for the original design is unchanged by the introduction of the blocking variable.

Values were not substituted for the Mean nor for the coefficient of the (AE) string; this latter remains represented by the value, 0.5(AE), or one-half the (AE) effect, multiplied by the appropriate plus or minus term, which in the case of Run 17, is +1.

The new mean is calculated using Run 20 Y', where the signs are all plus. Substituting in the following equation:

$$\text{Mean} + 0.5(AE)(+1) = \text{Run } Y'_{20}$$

thus

$$\text{Mean} + 2.3 = 23.25$$

or

$$\text{Mean} = 23.25 - 2.3 = 20.95.$$

Column Y" in Table 4 is obtained by subtracting the Mean value from each value in Column Y'.

The values in Column Y" can be used to estimate the effects of the individual interactions, either by summing them according to the sign matrix for each interaction or using Yates' algorithm (see Simon, 1977a). The sign pattern for the interactions shows that they are ordered according to Yates' standard order. Thus the calculations with Yates' algorithm would look like this with the values in Column Y":

Run #	Y" Data	Step #1	Effects-Sum	Effects	Effect Name
17	3.2	2.2	1.2	.6	AE
18	-1.0	-1.0	1.4	.7	BF
19	-3.3	-4.2	-3.2	-1.6	DG
20	2.3	5.6	9.8	4.9	CH

It is apparent that interaction CH accounted for most of the observed effect of the string and that interactions AE and BF are probably trivial.

34

It is difficult to conceive of a data collection plan
that can't be orthogonally blocked to some extent since
pairs of observations can be selected to be some fraction --
however small -- that is an orthogonal block to the original
set of data. Dykstra (1966, p 279) suggested that the
coordinates of each pair of new observation points should
equal the average of the original block.

SECTION III

USING YATES' ALGORITHM WITH SCREENING DESIGNS

Yates' algorithm (Simon, 1977a, pp 66-71; Davies, 1967, pp 263-265; Box, Hunter, and Hunter, 1978, pp 323-324; Yates, (1937) provides a convenient way of analyzing 2^k and 2^{k-p} designs with or without the use of a computer. With screening and other fractional factorial designs, the use of this algorithm is complicated because of the requirement to list conditions and the effects in Yates' standard order since with these designs all conditions of the factorial are not used and the experimental effects are aliased in sets. This complication is even more evident in the case of Simon's screening designs robust to trends (Simon, 1977a, pp 13-24). How Yates' algorithm is used to analyze a trend-robust screening design is explained below.

DETERMINING THE "STANDARD ORDER" OF THE EXPERIMENTAL CONDITION

In Table 5, an example of Simon's trend-robust screening design is shown, i.e., a Resolution IV 2^{8-4} design. To use Yates' algorithm, the experimental conditions must be listed in the "Standard Order": (1), a, b, ab, c, ac, bc, abc, d, and so forth. The conditions in Table 7, however, are aefg, bcdh, bcfg, and so forth. How do we reconcile the two lists and place the conditions used in the experiment in the standard order?

This is accomplished by ignoring the new screening design labels and corresponding names for the experimental conditions (e.g., aefg, bcdh,... (1), abcdefgh) and think in terms of the original factor labels. Remember, every fractional factorial has the same sign matrix as a complete factorial for fewer factors. Thus this 2^{8-4} design has the same sign matrix as a 2^4 factorial, although in this case the columns have been rearranged. To find the names of the conditions for the original design, we first find columns representing original factors A, B, C, and D. These are easy to identify by the alternating -+ pattern for A, --++ pattern for B, ----++++ pattern for C, and so forth. These columns are used to reconstruct the original names of the conditions. For example, take new experimental conditions, aefg, the first row of the matrix. Look for for the signs in that row associated with original factors A, B, C, D, which in this design are -, -, -, -, respectively. This new aefg is therefore the original (1) condition. Take row two with the new condition label, bcdh, and find the signs in that row for original factors A, B, C, D. This time the signs are +, -, -, - which is original experimental condition a. This process of finding the original condition labels from the signs associated with the original factor labels would continue until complete. (Note: While in this example, the conditions turn out to be in the standard order as listed,

TABLE 5. 2^{8-4} TREND RESISTANT SCREENING DESIGN

TEST ORDER	EXPERIMENTAL CONDITION	NEW SCREENING DESIGN LABELS* (Main Effects)*									(Two-Factor Interaction Strings)**						
		(I)	A	B	C	D	E	F	G	H	AH	AG	AF	AE	AD	AC	AB
1	AEFG	+	+	-	-	-	+	+	+	-	-	+	+	+	-	-	-
2	BCDH	+	-	+	+	+	-	-	-	+	-	+	+	+	-	-	-
3	BCFG	+	-	+	+	-	-	+	+	-	+	-	-	+	+	-	-
4	ADEH	+	+	-	-	+	+	-	-	+	+	-	-	+	+	-	-
5	BDFG	+	-	+	-	+	+	-	+	-	+	-	+	-	-	+	-
6	ACFH	+	+	-	+	-	-	+	-	+	+	-	+	-	-	+	-
7	ACDG	+	+	-	+	+	-	-	+	-	-	+	-	-	+	+	-
8	BEFH	+	-	+	-	-	+	+	-	+	-	+	-	-	+	+	-
9	CDEF	+	-	-	+	+	+	+	-	-	+	+	-	-	-	-	+
10	ABGH	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-	+
11	ABDF	+	+	+	-	-	-	+	-	-	-	-	+	-	+	-	+
12	CEGH	+	-	-	+	-	+	-	+	+	-	-	+	-	+	-	+
13	ABCE	+	+	+	+	-	+	-	-	-	-	-	-	+	-	+	+
14	DFGH	+	-	-	-	+	-	+	+	+	-	-	-	+	-	+	+
15	(I)	+	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+
16	ABCDEFGH	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ORIGINAL FACTORIAL LABELS		(I)	ABCD	ABC	ABD	ACD	AB	AC	AD	A	BCD	BC	BD	CD	B	C	D

*Three-factor interaction strings aliased with main effects.
**Two-factor interaction strings aliased with other two-factor interactions.

---

this need not always be the case and the investigator must determine it for each fractional factorial.)

APPLYING YATES' ANALYSIS TO THE DATA

In accordance with Yates' algorithm, the performance scores associated with each original label condition are listed in the standard order and the analysis is performed. Each effect or coefficient derived from the analysis is then identified, i.e., Mean, A, B, AB, C, AC, BC, ABC, D, etc. by listing the original factor labels in standard order.

SUBSTITUTING NEW EFFECT LABELS FOR ORIGINAL EFFECT LABELS

We relate the new labels in the Table 5 design to the original labels simply by replacement. In our example, we replace original factor label A with new factor label H, replace B with AD, AB with E, C with AC, AC with F and so forth.

SECTION IV

ANALYZING RESIDUALS

When economical multifactor designs are used, the investigator needs all the help he can muster to make certain that he is interpreting the data properly. **Residual analysis** is useful for this purpose.

Residuals are the unexplained variance in experimental data. A residual is the difference between an observed and a predicted score, the prediction being obtained from the regression equation derived from the data itself. Ideally, residuals have a zero mean, a normal distribution, a constant variance, and are independent of one another. Too much deviation from these ideals can distort the true interpretation of results based on analyses in which these conditions are assumed to be true. Inspection of the residuals can help the experimenter evaluate and interpret his data and decide what the next step in the analysis should be. Techniques for analyzing residuals have been discussed by Anscombe and Tukey (1963), Box and Cox (1964), Draper and Smith (1968), Daniel and Wood (1971), Wood (1973), Daniel (1976), and Box, Hunter, and Hunter (1978).

CALCULATING RESIDUALS

While predicted values may be determined individually by the direct use of the regression equation, use of the reverse Yates' algorithm is perhaps the quickest way to perform this task when a 2^k or 2^{k-p} experimental design has been employed. This technique has been described by numerous authors (Box, Hunter, and Hunter, 1978, p 344; Daniel, 1976, p 19; Hunter, 1966, Simon, 1977a, pp 78-79). Each predicted performance value is then subtracted from the observed performance value of each corresponding experimental conditions to obtain the residuals.

APPLICATIONS OF RESIDUAL ANALYSIS

Residual analysis may be used to:

1. Check how well the data meets the assumptions of normality and zero mean. A frequency plot of the residuals or an ordered plot on normal probability paper (see item 4 below) will indicate how close the residuals are to being normally distributed with zero mean. Since finite data will show variations from the ideal, some experience is needed to interpret these plots. Daniel and Wood (1971, Appendix 3R) and Daniel (1976, Appendix 6A) illustrate how such fluctuations might appear if due to chance alone.

2. Check how well the data meets the assumptions of homogeneity. A plot of the residuals against the estimated values or the independent variables can reveal when the variance is not constant, being larger for some variables than others. When there are too many independent variables, graphic plots will be difficult to interpret unless subsets of variables are examined.

3. Calculate the precision of the fitted estimate. The residual variance is considered to be the "error" variance of both the ANOVA and regression models. The residual variance is calculated by dividing the residual sum of squares, $(Y_i - \hat{Y}_i)^2$, by $(N-K-1)$, where Y_i is the observed response and $\hat{Y}_i$ is the predicted response at condition i, N is the total number of observations, and K is the number of factors being isolated. The standard deviation is the square root of that value; it can be used to calculate confidence limits. The standard error of the statistic is the standard deviation divided by the square root of $(N-K-2)$. If the ratio between the residual sum of squares and total sum of squares is subtracted from one, the result equals the multiple correlation squared, or the proportion of variance accounted for by the terms of the regression equation.

4. Detect outliers and errors in data collecting, recording, and/or calculating. Ideally residuals should be distributed normally with a mean of zero. In an ordered plot on normal probability paper, the i-th value from the bottom, z_i, is plotted against a value, $a_{i/n}$*, chosen to be typical of the i-th value from the bottom in a sample of n from a unit normal distribution. When the plot fails to follow a straight line, deviant individual measures may suggest that something unusual has occured and should be explained.

---

*The value of $a_{i/n}$ represents the probability point on the normal probability scale for the i-th position of n data points. It has been calculated in different ways. Daniel (1978) used the equation

$$a_{i/n} = (i - 0.5)/n$$

While Anscombe and Tukey (1963, p 145) suggest that

$$a_{i/n} = (3i - 1)/(3n = 1)$$

is a "shade better" calculation. For n = 15, this would locate the 12th point at the .767 probability point using Daniel's calculation and .761 using Anscombe and Tukey's.

Anscombe and Tukey propose plotting a variation
on the usual graphic plot of residuals, in order
to discover outliers and errors. They write (p 145):
"If $\acute{z}$ represents the median of the z's, plotted at
$(0, \acute{z})$, then the slope of the secant *leading from*
$(0, z)$ to $(a_{i/n}, z_i)$ is

$$\frac{z_i = \acute{z}}{a_{i/n}}$$

and may usefully be calculated for i's sufficiently far
from the median. (Omitting the middle third of the i's
seems satisfactory.) An exactly straight line for the
cumulative on normal probability paper corresponds
to exactly constant values for these slopes. A plot
of $(z_i - \acute{z})/a_{i/n}$ against i is thus quite revealing..."
By way of illustration, they suggest that the plot
would show the following characteristics for each
specific perturbation of the data:

DATA	PLOT
• Single outlier	Large value for i = 1 or n, depending on sign of outlier
• Number of outliers of each sign, or Symmetrical distribution with more extended tail than normal	Plots turn up at both ends
• Tendency to skewdness	Plot higher at one end than the other

The presence of an outlier may mean that an error
has occurred or it may mean that an unusual but
valid event has occurred. At best, the residual
analysis alerts the investigator and encourages
him to search further for an explanation. Anscombe
and Tukey warn that with residual analysis it is
relatively easy for one kind of misbehavior of the
data to simulate another. For example, they note
(p 143) that "a single very wild value will (i) act
like an outlier, (ii) lead to very non-normal values
of estimated shape coefficients, (iii) indicate
enhanced variability for each of the subgroups to

which it belongs, (iv) appear to constitute a measurable dependence of variability of response upon level of response." They also emphasize the ease with which the strength of the evidence offered by a single plot may be overrated and they encourage the search for repeated occurances of a residual pattern from several bodies of data.

Anscombe and Tukey also point out the dangers of applying numerical procedures to data before outliers are properly dealt with. Removal may be the best way to handle erroneous data, either analyzing the remaining data with the values missing or assigning modified values in their places. The circumstances and the purpose of the experiment help determine the best way to handle outliers.

5. Help decide when to stop adding terms in a cumulative model. The results from an unreplicated screening design may be ordered according to the magnitude of the effects of each source of variance and a cumulative proportion of variance can be assigned to each succeeding value (Simon, 1977a, pp 75-83; 1979, p 41). The investigator must decide where to draw the line between effects that probably are critical and those that are not. If residuals are obtained using the subset polynomial and no unusual patterns are observed and a test of the lack of fit finds that the model adequately represents the response surface, the investigator may stop including terms beyond that point.

6. Indicate the adequacy of the model. If "pure" error, as obtained from repeated measures at the center of a central-composite design, is isolated from the residual variance, the remaining variance represents the degree to which the estimated model fails to fit the data. When residuals are plotted against the predicted values, the presence of a linear or quadratic trend suggests that a data transformation is needed or that additional terms must be added to the model.

7. Identify terms contributing to large interaction effects. Valid three-factor interactions are infrequent in human performance data (Simon, 1977a). Valid four-factor or higher interactions are so unlikely that if their effects are large, they should be suspect. Quite often a large higher-order interaction may indicate the presence of an aberrant observation. An examination of the residuals when a questionable term is excluded from the fitted model will often indicate which conditions are responsible.

Hunter (1966) gives an example to show how residual analysis may facilitate the interpretation of such an interaction. In a 2^4 factorial design, two of the factors, their two-factor interaction, and the four-factor interaction showed large effects. The inverse Yates' algorithm was used to determine the predicted performance when the model included only the coefficients of the two-factors and their interaction with all other coefficients set equal to zero, including the non-trivial one associated with the four-factor interaction effect. When the residuals were obtained for each term of the 2^4 factorial, one condition was found to show an unusually large response when compared with the estimated standard deviation for the data.* It was discovered that the large response was due to a recording error. When it was corrected, the aberrant four-factor interaction disappeared.

8. **Reveal distortions introduced by external sources of variance.** Residuals should be uncorrelated with any external source of variance. If, for example, residuals are plotted as a function of time, we do not expect to find that they increase or decrease in magnitude, become more or less variable, or even show curvilinear relationships over time. The presence of any of these patterns suggests that *the data be transformed* to eliminate the effect or that other terms be added to the model to account for time effects.

If the investigator is concerned that his experimental data may be distorted by uncontrolled and unplanned changes that took place in the environment when the data was being collected, he may compare the residuals associated with the different conditions to see if they show corresponding changes.

Daniel and Wood (1971, p 32) state that plotting techniques can not be expected to work well with less than 20 observations, and are more efficient when the number exceeds fifty. Still, an investigator may use them to enhance the effectiveness of his visual inspection of the data prior to formal analysis, provided he does so with restraint.

---

*In this example, the standard deviation was obtained by combining the sum of squares for the eleven terms of the factorial that were considered trivial (i.e., all except the two factors, their interaction, and the four-factor interaction) and dividing by N-K-1 (or 11 in this example) and taking the square root of the results.

SECTION V

IDENTIFYING THE EXPERIMENTAL CONDITIONS IN 2^{k-p} DESIGNS WHEN GIVEN THE DEFINING GENERATORS

At times, 2^{k-p} fractional factorial designs are described only in terms of the defining generators.* An investigator may wish to know what experimental conditions make up the completed design. This information is needed to run an experiment or if a design is to be fractionated, to identify the conditions to be used without needing to construct a complete sign matrix.

To illustrate this process, let us begin with the following set of defining generators:

I, ABC, ADE.

Generators are independent of one another and <u>cannot</u> be obtained by multiplying together other generators. <u>Contrasts</u> are obtained by multiplying ever factorial combination of the defining generators by one anoth . In this case, with two generators, there is only one additional combination. The complete set of defining contrasts here would be:

I = ABC = ADE = BCDE.

This set of defining contrasts describes a quarter replicate of a 2^5 design containing five factors, A, B, C, D, and E. The design requires $2^{5-2} = 2^3 = 8$ experimental conditions. It is a Resolution III design, in which all main effects can be isolated from one another but not from two-factor and higher-order interactions.

To form different fractions (i.e., blocks of the total factorial), the design generators may be assigned a positive or a negative sign. The signs of the complete set of defining contrasts will follow the usual arithmetic rules for multiplying values with the same or different signs.** Thus:

---

*Readers who are unfamiliar with the terms or basic mechanics discussed in this paper are referred to Box and Hunter (1961), Davies (1967) or Simon (1973).

**Of course, we aren't really multiplying signs; we are multiplying coefficients of plus or minus one. We drop the one to make the tables less confusing.

	Generators			Remaining Contrast
	I	ABC	ADE	BCDE
1)	+	+	+	+ ◄—(which is the product of two plus signs)
2)	+	+	-	- ◄—(which is the product of a plus & minus sign)
Blocks 3)	+	-	+	-
4)	+	-	-	+

These contrasts are actually read as a string of effects all aliased with the Identity

Block 1 =   (I + ABC + ADE + BCDE)

Block 2 =   (I + ABC - ADE - BCDE)

and so forth.

From these we can easily determine the aliases for each effect in any block. For example, in Block I:

(A + BC + DE + ABCDE) ◄—(The product of A
(B + AC + ABDE + CDE)   and the identity
(AB + C + BDE + ACDE)   string for Block 1)

and so forth. For Block 2, the string of aliases for B would be, for example:

(B + AC - ABDE - CDE).

To find the experimental conditions in a particular block, we must assign signs to each factor (in this case, to A, B, C, D, and E) in combinations that produce the correct signs for the contrasts given for the particular block. Let us use Block 3 to illustrate this, where

Block 3 =   (I - ABC + ADE - BCDE).

Prepare at least one column for each factor. Sometimes it helps, but is not necessary and may even be awkward as the number of defining contrasts increases, to group them as they appear in the contrasts. Thus we might head each column as:

1)   A   B   C   D   E

or   2)   ⌊A   B   C⌋   ⌊A   D   E⌋   ⌊B   C   D   E⌋

or   3)   ⌊B   C   A⌋   D   E   ABC
                            ⌐   ADE
          └─►   ◄─┘   BCDE

To illustrate how we build the sign matrix, let us use the second layout where each contrast has its own columns. We must develop eight independent experimental conditions. The steps below show how the conditions in Table 6 were derived.*

Starting with contrast ABC, which for Block 3 has a minus sign associated with it, we assign signs to A, B, and C individually so that their product will always equal minus. This means that there must be an unequal number of minus signs assigned, i.e., either one or three. The four possible combinations for A, B, and C, that meet this criterion are -++, +-+, ++-, and ---. Next we assign signs to A, D, and E to create a positive ADE contrast. The available selections are restricted by the signs already assigned to A. For this contrast, since ADE is assigned a plus sign in this Block, there must be an equal number (i.e., two) of minus signs in every combination. Thus, if A is already minus, then there are two possible combinations for D and E. One would be +, - respectively and the other would be -, +. When A is plus, then D and E must be either ++ or --. Having filled these two contrasts the columns for B, C, D, and E are already determined. If the first two contrasts have been done correctly, the third, BCDE with a minus sign, will also be correct, with an odd number of minuses (1 or 3) each time. The combined signs are shown in Table 6.

TABLE 6. BUILDING THE SIGN MATRIX & IDENTIFYING CONDITIONS

Factor Headings (Grouped by Contrast)

	-(A B C)	+(A D E	-(B C D E)	Block 3 Conditions
1)	- + +	- + -	+ + + -	bcd 1)
2)	- + +	- - +	+ + - +	bce 2)
3)	+ - +	+ + +	- + + +	acde 3)
4)	+ - +	+ - -	- + - -	ac 4)
5)	+ + -	+ + +	+ - + +	abde 5)
6)	+ + -	+ - -	+ - - -	ab 6)
7)	- - -	- + -	- - + -	d 7)
8)	- - -	- - +	- - - +	e 8)

---

*It is only necessary to assign the signs to all of the defining generators. When this is done properly, the remaining defining contrasts will be correct. When the fractional factorial is small, writing out all of the defining contrasts could become burdensome for this purpose.

The final step is to use the sign matrix to name the experimental conditions that make up the fractional factorial for this particular Block 3. We do this, of course, a row at a time by citing the letters of the factors with a plus sign. Thus within the boxes surrounding the five factors in the table, we find on the first row, A, B, C, D, and E have the signs -, +, +, +, -, respectively. Since Factors B, C, and D were given plus signs in that row we name the first experimental condition bcd. The names of the other conditions are shown in the right-hand column of the table.

SECTION VI

AN ECONOMICAL DESIGN FOR SCREENING INTERACTION EFFECTS

DeGray (1968) proposes an in eresting economical, multi-factor design for identifying a certain type of interaction response during the preliminary and screening phases of a research program. This type of interaction is characterized by a pattern of responses in which nothing of consequence happens as each experimental condition is tested until a particular combination of factor levels are combined; then a major response occurs. He refers to this type of interaction as "coactive."

DeGray gives the following example to show the effect to be expected. A spark is passed through four chemical atmospheres composed of the factorial combinations of the presence or absence of two chemicals, hydrogen and oxygen, i.e.,

(1)	None
a	Hydrogen only
b	Oxygen only
ab	Mixture of hydrogen and oxygen.

A coactive interaction is in icated when nothing happens in the first three tests and an explosion occurs in the last. If the magnitudes of these responses are designated 0, 0, 0, and 1000 respectively, an analysis of variance would reveal the effects of the three sources of variance to be

A	500
B	500
AB	500

When the effect of the interaction term is approximately the same size as the main effects, as in this example, the presence of a "coactive" interaction is indicated. The results from the analysis of variance tends to cloud the true implications of the data.

APPLICATIONS TO EQUIPMENT DESIGN PROBLEMS

The design might be used when the investigator expects that only one combination of equipment parameters in a 2x2 matrix will show an exceptionally strong effect. This form of interaction is referred to as "boom" effect. For example, in the design of pilot-training simulators, one needs to know how many and which of the six degrees of motion are necessary for effective training. Several

approaches can be considered. One might study the main effects of the six degrees of motion in a screening design. With 16 observations, all main effects can be isolated from the effects of strings of two-factor interactions but remain aliased with three-factor interactions. For this particular problem, however, one might suspect that it would be more informative to obtain and understand the effects of combinations of degrees of motion, i.e., the interactions, rather than the main effects. If so the conventional screening design would be the worst data collection plan to use. On the other hand, it would not be economical to do a complete factorial design to look at all the interactions. With six degrees of motion being present or absent, this would require the investigation of 64 combinations. A less systematic approach has sometimes been proposed in which only particular combinations, selected on the basis of special knowledge, might be studied but this is always subject to the dangers of omission, particularly at the early phases of a research program. As an alternative to these approaches, DeGray's oactive designs might be employed for a preliminary economical look at the problem.

CONSTRUCTING A COACTIVE DESIGN*

The design requires a minimum of N combinations to identify the effective interactive (or coactive) combinations of N variables. These N combinations consist of N arrangements of the N variables taken (N-1) at a time. Starting with all factors present (or at a high level), the remaining combinations are made up by dropping one factor each time (or by setting it at a low condition). DeGray felt that it was unnecessary to include the all-high combination; however, in behavioral studies, both all-high and all-low (or absent) combinations might be included to provide a frame of reference against which the other combinations could be compared.

Thus a design for examining interactive effects among six-degrees of motion designated A through F, would look like this:

(1)	A	B	C	D	E
(2)	A	B	C	D		F
(3)	A	B	C		E	F
(4)	A	B		D	E	F
(5)	A		C	D	E	F
(6)		B	C	D	E	F

---

*DeGray discusses why this design differs from the classical one-factor-at-a-time, main effect design.

where the presence or absence of a letter indicates the presence or absence of the corresponding degree of motion. As suggested earlier, we might also wish to include two other combinations, ABCDEF and (1).

INTERPRETATION OF THE DATA

The data would not be combined arithmetrically. Instead, the pattern of responses would be examined. For example, if all of the combinations gave a positive response except No. 3, then it would be Variable D that probably caused the reaction. If combinations Numbers 4 and 6 both failed to give an adequate response, the interaction between A and C would be suspected of causing the reaction. If Numbers 3, 4, and 6 failed to react, then interaction ACD would be suspected. On the other hand, if combinations of variables acted as inhibitors, the pattern of responses would be interpreted in reverse. The investigator should be able to determine which of the opposing situations exists, either by using knowledge he already has or by making a few additional observations.

DeGray points out that the design works best when a positive response is definitely positive and a negative response is definitely negative. This may not be typical of behavioral data, and certainly not with the clarity that might be found in certain chemical reactions. The design still can be used if an investigator suspects certain interactions might be important and inspects the data accordingly. If used cautiously, this might be the most economical means of unraveling a great many two-factor interactions effects confounded in strings in the Resolution IV screening design.

Interpreting the results from this design can be affected if sequence effects are operating. Simon (1974) has suggested ways that these might be minimized during the data collection phase. Furthermore, in our degrees of motion example, no consideration was given to the fact that other equipment variables may be operating and could interact with the other degrees of motion. These might be held constant in an exploratory study or included in the coactive design if the investigator deems it necessary.

It should be obvious that the investigator must take the same precautions with this as with any other experimental design, whether exploratory or primary. The technique is worth trying and gives a systematic economical approach to questions that otherwise could be costly to answer.*

---

*Cotter (1979), without referencing DeGray's work, presents an expanded -- 2n + 2 -- design composed of the n conditions in DeGray's coactive design, plus n "foldover" conditions, plus one condition with all factors at the high level and one with all at the low level. Though slightly better than DeGray's, the precision of this design is still poor.

SECTION VII

GRAPHIC METHOD AND INTERNAL COMPARISON
FOR MULTIPLE RESPONSE DATA

The 2^{k-p} screening designs described by Simon (1973, 1977a) do not rely on replication to provide the error variance needed in a test of statistical significance of the effects. Daniel (1959, 1976; also see Simon, 1977a, pp 38-98) proposed the use of order statistics and internal comparison procedures to detect those uniresponse effects that were probably not due to chance. Internal comparisons procedures permit simultaneous comparisons among effects with the aid of a statistical standard which is, at least in part, generated internally by the data. The process is conceptually simple. If the univariate effects (or contrasts) determined from the analysis of the fractional factorial (screening design) are due to chance, when ordered and plotted properly on normal probability paper* they will lie along a straight line. Those that are larger than might be expected by chance will deviate noticeably from the line.

Wilk and Gnanadesikan (1964; also see Roy, Gnanadesikan, and Srivasta, 1971, Chapter VIII; also see Simon, 1977a, pp 151-158 for a general description of the procedure) describe how to use a similar graphic method to evaluate the effects from an unreplicated design when there are multiple responses. While there are times when the purpose of the mission (and not the statistics) determines the relative weights to be assigned to multiple criteria, there will be circumstances when the investigator may wish to rely on simple statistical combinations that take into consideration the correlated relationships. Since the procedure proposed by Wilk and Gnanadesikan is rather involved and is explained in matrix algebra terms, it is described below along with an example to facilitate its use.

THE GENERAL APPROACH

The procedure in some respects is analogous to Daniel's wherein the multivariate responses of the 2^{k-p} effects are tested simultaneously by means of graphical internal comparisons. In this approach, we will need a multivariate estimate of the effect and suitable graph paper on which to plot it. The multivariate estimate will be the squared Euclidean distance between the centroids (of high and low conditions) in the multivariate space. These will be plotted on the appropriate quantiles of the standard gamma distribution. If the null hypothesis is correct, the plots

----

* In the original article, Daniel (1959) proposed using half-normal probability paper on which to plot the data. Later, in his 1976 book, he suggested that more information would be revealed if full normal probability paper were used.

will lie along a straight line. If it is not correct, the largest effects will appear to curve away from the straight line configuration. This procedure is somewhat more complicated to apply than is the plotting of univariate effects on normal probability paper since two parameters, λ and η, of the gamma distribution must be estimated.

Let us illustrate the procedure step by step, using a numerical example prepared by Weinman (1979).

THE WORKING DATA

The fictitious data for this example is shown in the first three columns of Table 7. In the first column are the 32 experimental conditions of a 2^5 factorial design listed in Yates' standard order. In the next two columns are the values for the two responses, y_1 and y_2, for each condition. The second variable, y_2 was actually created by adding random digits, $(\pm0, \pm1, \ldots \pm9)$, to yl.

OBTAINING THE SQUARED DISTANCES

The next two columns, Y_1 and Y_2, give the effects (mean differences or contrasts) for the values in y_1 and y_2, respectively. The effects are obtained by applying Yates' algorithm to the columns for y_1 and y_2 separately. Since the effects are arranged in standard order, the sources of their identities are composed of the same letters as the conditions on the same line.

To calculate the squared distances for the Identity Matrix, sum $(Y_1^2 + Y_2^2)$. For example, using the Y_1 and Y_2 values for the S effect in Table 7, i.e., -4 and -5, the squared distance for S is $(-4)^2 + (-5)^2 = 41$. These are listed in Table 7 in column six, and the source of each is the same as for the individual effects. If there had been more than two responses, the squared distances for the Identity column would have been $(Y_1^2 + Y_2^2 + \ldots + Y_p^2)$. This calculation assumes that the relationship among the responses is a simple linear one, that is, they are weighted equally without standardization.

DETERMINING THE PARAMETERS OF THE GAMMA PLOT

First determine how many of the K squared distances are likely to be trivial. Ordinarily the higher-order interactions would be selected. In our example, all three-factor interactions and higher --16 distances (Table 7, column 1) - were chosen. Of these, the M smallest squared distances are selected so that the ratio K/M is at least $3/2 = 1.5$, and preferably slightly larger. Avoid any potentially critical distances. In our example, the eight distances of the Identity matrix (Table 7, column 6) with

TABLE 7.   RAW DATA, EFFECTS, AND SQUARED DISTANCE

Experimental Conditions	Responses		Effects[#]		Squared Distances	
	Y_1	Y_2	Y_1	Y_2	Identity	S^{-1}
1	2	3	4	5	6	7
(1)	3.5	7.0	-0.06.	0.094		
s	11.5	-15.5	-4.000	-5.000	41.0	2.34
d	8.0	10.0	7.87	8.312	131.1	8.36
sd	-2.0	-2.0	2.5.	2.875	14.5	0.82
n	4.5	9.0	1.56.	1.938	6.2	0.35
sn	-18.0	-19.0	1.68.	2.250	7.9	0.50
dn	4.0	6.5	2.562	2.688	13.8	0.89
sdn*	5.5	3.5	0.938	1.000	1.9*	0.12*
p	-1.0	0.5	-2.625	-3.875	21.9	1.74
sp	-14.5	-18.0	1.500	2.688	9.5	1.27
dp	9.0	5.5	-0.250	0.125	0.1	0.17
sdp*	-4.5	-9.0	-5.875	-7.062	84.4	3.45
np	-11.5	-15.5	-2.562	-1.875	10.1	1.82
snp*	-6.5	-6.5	0.562	1.438	2.4*	0.67*
dnp*	1.0	3.0	0.562	3.250	10.9	6.68
sdnp*	3.0	3.5	-0.312	-1.438	2.2*	1._6
k	2.5	0.5	3.750	4.812	37.2	2.19
sk	-10.0	-9.0	4.250	6.125	55.6	4.14
dk	-4.0	0	-2.000	-1.562	6.4	0.97
sdk*	-7.0	10.0	-0.750	-1.000	1.6*	0.10*
nk	6.0	10.5	2.188	1.188	6.2	2.18
snk*	0.5	0	-3.062	-3.500	21.6	1.23
dnk*	7.5	3.5	1.188	0.438	1.6*	0.96
sdnk*	16.5	17.5	-1.094	0.750	1.7*	3.94
pk	-4.5	-4.5	-0.250	0.625	0.4	0.80
spk*	7.5	9.5	-1.750	-2.562	9.6	0.75*
dpk*	2.5	3.0	-0.8	-0.625	1.2*	0.22*
sdpk*	-1.5	-2.0	-1.875	-2.062	7.8	0.46*
npk*	-5.5	-8.5	-2.43.	-2.500	12.2	0.83*
snpk*	-5.5	-5.5	-3.188	-3.438	22.0	1.35
dnpk*	9.0	11.0	1.438	1.875	5.6*	0.34*
sdnpk*	1.0	4.0	2.438	1.938	9.7	1.39

[#]The sources for the effects (and the squared distances) are labelled with capital letters that correspond to the letters of the Experimental Condition on the same row. For example, Factor S has effect -4 for Y_1 and -5 for Y_2.

an asterisk beside them were chosen*. This made the K/M
ratio equal 16/8 = 2, an acceptable value.

The M values just selected will be used to obtain an
estimate of $\hat{\eta}$, the shape of the gamma distribution. Two
parameters, P and S, are needed.

P = geometric mean of the M distances divided
    by the largest of the M values. The geo-
    metric mean is equal to the product of the
    M distances raised to 1/M.

S = arithmetric mean of the M distances divided
    by the largest of the M values. The arith-
    metric mean equals the sum of the M distances
    divided by M.

In our example, P = 2.034/5.6 = 0.36, and S = 2.275/5.6
= 0.41. To find Eta = $\hat{\eta}$, use Table IV, beginning on page
198, in Roy, Gnanadesikan, and Srivastava's book (1971).**
Find the table for appropriate K/M ratio and then look up
P and S. It may be necessary to make a bilinear interpola-
tion to properly represent your values if they are located
between those listed in the tables.

A bilinear interpolation can be made using the following
equation:

$$\hat{\eta} = \eta_{11} (1 - a - b + ab) + \eta_{21} (a - ab) + \eta_{12} (b - ab)$$

$$+ \eta_{22} (ab)$$

where

$$a = \frac{P - P1}{P2 - P1} \quad \text{and} \quad b = \frac{S - S1}{S2 - S1}$$

and P and S are the values you obtained, and P1, P2, S1, and
S2 are the two pairs of values in the tables that bracket P
and S. η_{ij} is the tabled value at P_i and S_j.

In our example, for K/M = 2.0, we find that P = .36 is
one of the values listed; we are lucky. Our S of 0.406,
however, is between the tabled values of .40 and .44, where

---

*In retrospect, one might question whether or not dis-
tance X_8 = 5.6 should have been included in the M set, since
it appears to be unduly large.

**These tables are also given in an article by Wilk,
Gnanadesikan, and Huyett (1962).

the ηs are 1.443 and 1.306, respectively. The calculations
for making the bilinear interpolations are shown in Table 8
for this data (although in this case, no interpolation was
needed for the P_i dimension. The $\hat{\eta}$ is estimated to be 1.4.

To find the quantiles of the gamma distribution, we need
Table VII beginning on page 208 in Roy, Gnanasedikan, and
Srivastava's book (1971)*. The percentages and associated
quantile values will be found in that table under the ETA
value just determined. In our example, under the ETA = 1.4,
we find the following values listed:

Percentage	Quantile
0.1	8.4321789E-03
0.5	2.6824095E-02
0.7	3.5971068E-02

on up to

99.5	6.2058056E 00
99.9	7.9005137E 00

With these pairs of values serving as coordinates, we plot
the function relating percentage to quantiles (see Figure 4).

---

*These tables can also be found in an article by Wilk,
Gnanadesikan, and Huyett (1962).

TABLE 8.   EXAMPLE OF BILINEAR INTERPOLATION

P1 = .36 = P2

	η_{11}	η_{21}
S_1 (.40)	1.443	1.443
	η_{12}	$_{22}$
S_2 (.44)	1.306	1.306

$$a = \frac{.36 - .36}{.36 - .36} = 0 \qquad\qquad b = \frac{.406 - .40}{.44 - .40} = .15$$

$$\hat{\eta} = \eta_{11}\ (1-a-b+ab) = \eta_{21}\ (a-ab) + \eta_{12}\ (b-ab) + \eta_{22}\ (ab)$$

$$\hat{\eta} = 1.443\ (1-0-.15+0) + 1.443\ (0-0) + 1.306\ (.15-0) + 1.306\ (0)$$

$$\ddot{\eta} = 1.443\ (.85) + 1.306\ (.15)$$

$$\hat{\eta} = 1.42 \text{ or } \underline{1.4}$$
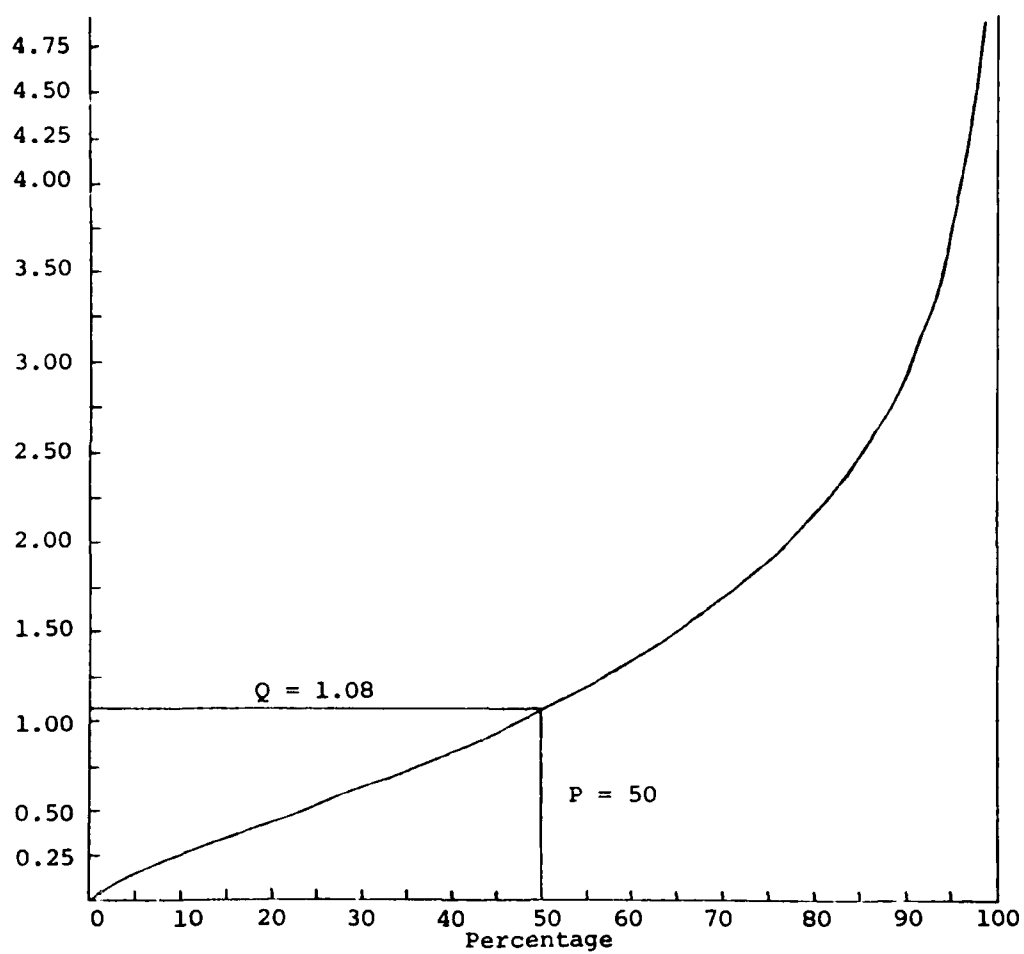
Figure 4. Plot of Quantiles Against Percentages
($\hat{\eta} = 1.4$)

THE PLOT

Table 9 is to be prepared. The first column consists of the ranks listed from 1 to N-1, in our case, 31. Next, the percentage, b_i, associated with each rank is calculated using the following equation:

$$b_i = \frac{i - 0.5}{L} \times 100$$

for rank i = 1, 2, ... L, where L = N-1, or 31 in our example. Later, if the investigator wishes to drop some of the largest distances, he would go through this same procedure, but would use the smaller value of L. These percentage values would be listed beside the appropriate rank as shown in Table 9, column `. Next we will list in column 3, the quantiles associated with each of the percents. These are determined from the function drawn in Figure 4. For example, b_{16} or ° , equals quantile x_{16} or 1.08. In column 4, we list the squared distances of the Identity matrix found in Table 7, but ordered from the smallest to the largest. In column 5, the source associated with each distance is given.

L points are next plotted on ordinary graph paper with the squared distances on the ordinate and the quantiles on the abscissa. The 31 points in our example are shown plotted in Figure 5. Although reasonable care should be taken in the plotting, for the smaller ranks, it may not be necessary to plot every point.

INTERPRETATION

Inspection of Figure 5 shows that the distances at the lower ranks tend to lie along a straight line*. At the higher ranks, certain distances begin to deviate above the

---

* Actually, a break is visible between the 9th and 10th point, creating two straight lines with approximately the same slope. This suggests that the error variance is not homogeneous. Weinman (1979) proposes a possible explanation for this based on the fact that the x_1 data had been taken from some published by Yates. He writes: "Yates mentions a ridge of fertility running through the field. This could account for such a jump. It's likely the hypothesis of homogeneous variances is not correct. I would also note that the field was laid out in four blocks and none of the 14 longest distances come from block three. This suggests that block three has a different effect on yields from the other three blocks, possibly because of the ridge of fertility."

TABLE 9. DATA REQUIRED TO PLOT THE ORDERED
DISTANCES FOR THE IDENTITY MATRIX

Ranks	Percentages	Quantiles	Ordered Distances	Sources
1	1.6	.06	0.1	DP
2	4.8	.14	0.4	PK
3	8.1	.21	1.2	DPK
4	11.3	.27	1.6	DNK
5	14.5	.33	1.6	SDK
6	17.7	.39	1.7	SDNK
7	21.0	.46	1.9	SDN
8	24.2	.52	2.2	SDNP
9	27.4	.58	2.4	SNP
10	30.6	.64	5.6	DNPK
11	33.9	.71	6.2	N
12	37.1	.78	6.2	NK
13	40.3	.85	6.4	DK
14	43.5	.92	7.8	SDPK
15	46.8	1.00	7.9	SN
16	50.0	1.08	9.5	SP
17	53.2	1.16	9.6	SPK
18	56.5	1.25	9.7	SDNPK
19	59.7	1.34	10.1	NP
20	62.9	1.45	10.9	DNP
21	66.1	1.55	12.2	NPK
22	69.4	1.68	13.8	DN
23	72.6	1.81	14.5	SD
24	75.8	1.95	21.6	SNK
25	79.0	2.12	21.9	P
26	82.3	2.32	22.0	SNPK
27	85.5	2.55	37.2	K
28	88.7	2.83	41.0	S
29	91.9	3.20	55.6	SK
30	95.2	3.75	84.4	SDP
31	98.4	4.82	131.1	D
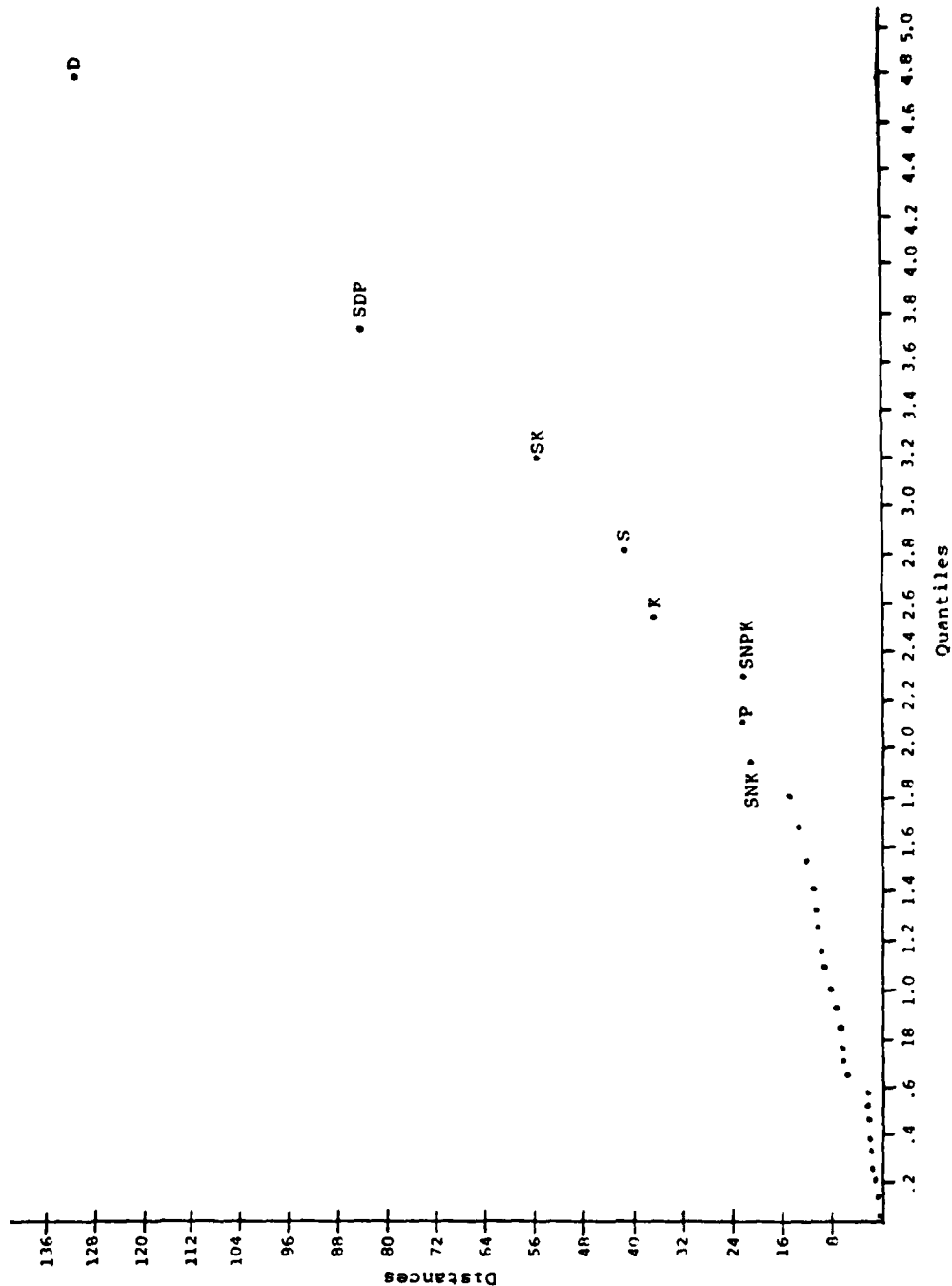
Figure 5. Plots of Ordered Distances Against
Quantiles for the Identity Matrix.
(L = 31, K = 16, M = 8, η = 1.4)

line extrapolated from the line formed by the points at the left. The interpretation process corresponds to that for uniresponse graphic plots, i.e., those distances lying off the line are larger than might be expected by chance. In this example, distances for D, SDP, SK, S, and K, are well above the line, and SNPK, P, and SNK are worthy of further investigation. One might repeat this entire process but dropping the first four distances, reducing K to 27. This procedure might be repeated several times for the remaining larger distances.

The graphic process provides one additional criterion to help the investigator interpret his data. The larger the number of effects, the more likely the idealistic principles will behave properly (providing the experimenter has done the rest of his job properly). Certainly the multivariate nature of these effects makes any interpretation more complex than would be the case with a uniresponse design. This complexity is further increased by the aliasing present in the fractional factorial designs. Only experience will overcome these difficulties. For the present, however, the investigator should check and double check his conclusions against a variety of criteria, using the analyses to assist him rather than to lead him. Gnanadesikan (1963) emphasizes that the use of a multiresponse analysis does not preclude examining each response singly to better understand the multiresponse data.

OTHER MULTIVARIATE MODELS

Most statistical analysis is based on a linear ordering of the data. This is a characteristic of real numbers but not of vectors or combinations of real numbers. Thus, the ordering of multivariate data involves the use of a metric or some measure of size*. Distance, d_i, does not have to be measured as "squared distance" as was done in the above example. Actually the complete expression should be:

$$d_i = x_i' A x_i, \qquad i = 1, 2, \ldots, K$$

where A is a compounding matrix, positive semi-definite, but otherwise selected at the discretion of the investigator.

The Identity matrix used in our illustration weights the variables Y_1 and Y_2 equally and does not take into account any correlation that might occur between the two. For that

---

* Barnett (1976) has written a comprehensive article describing the problems and proposed solutions for the ordering multivariate data.

model, the compounding matrix A is a unit matrix, which has ones on the principal diagonal and zeros off the diagonal.

A slightly more complex matrix would be a diagonal matrix with reciprocals of the variance of Y_1 and Y_2 on the diagonal. This is equivalent to defining "distance" as

$$D = \frac{Y_1^2}{S_1^2} + \frac{Y_2^2}{S_2^2}$$

The variables would be "adjusted for size," but would still be treated separately with their effects simply added together.

THE S^{-1} MATRIX

To take into account the covariation among responses, a non-diagonal matrix is required. A standard matrix used for this purpose in multivariate analysis is S^{-1}, where S is a matrix of estimates of variances and covariances of the variables. That is,

$$S = \begin{pmatrix} S_1^2 & S_{12} & S_{13} \cdots S_{1k} \\ S_{21} & S_2^2 & S_{23} \cdots S_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ S_{k1} & S_{k2} & S_{k3} & S_k^2 \end{pmatrix}$$

where the diagonal elements, S_i^2, are the sample variances of the variables and the off-diagonal elements, S_{ij}, are sample covariances between variables i and j. That is,

$$S_i^2 = \frac{N \Sigma Y_i^2 - (\Sigma Y_i)^2}{(N-1)}$$

and

$$S_{ij} = \frac{\Sigma Y_i Y_j - \frac{(\Sigma Y_i)(\Sigma Y_j)}{N}}{(N-1)}$$

In general, one would use a computer to both calculate the S matrix and to invert it (i.e., find S^{-1}). Computer packages for these purposes are readily available.

In the original example, with only two responses, the S matrix is only 2 X 2 and inversion is simple. If

$$S = \begin{pmatrix} a & b \\ b & d \end{pmatrix}$$

then

$$\bar{S}^1 = \begin{pmatrix} \dfrac{d}{ad-b^2} & \dfrac{-b}{ad-b^2} \\ \dfrac{-b}{ad-b^2} & \dfrac{a}{ad-b^2} \end{pmatrix}$$

The rule is simple: interchange d and a, put minus signs on the b's, and divide every element by $(ad - b^2)$.

Continuing with the original example, if we can use this approach on columns IV and V of Table 7, $S_1^2 = 7.76$, $S_2^2 = 10.82$, and $S_{12} = 8.71$, so

$$S = \begin{pmatrix} 7.76 & 8.71 \\ 8.71 & 10.82 \end{pmatrix}$$

and

$$S^{-1} = \begin{pmatrix} 1.349 & -1.086 \\ -1.086 & 0.986 \end{pmatrix}$$

The distances, D, are defined as

$$D = (Y_1 \ Y_2) \ S^{-1} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

or in our example,

$$D = \begin{pmatrix} Y_1 & Y_2 \end{pmatrix} \begin{pmatrix} 1.349 & -1.086 \\ -1.086 & 0.968 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

$$= 1.349 \ Y_1^2 - 2(1.086)Y_1Y_2 + 0.968 \ Y_2^2$$

The D's calculated in this manner are shown in the last column of Table 7.

The procedure from this point on, given these distances based on the S^{-1} matrix rather than the Identity matrix would be the same. When L = 31, K = 16, and M = 8, the eight smallest distances (among the 3-factor or higher interactions are marked with asterisks in the last column). Note that five of the smallest are the same as with the Identity matrix, but three are not. For those readers who wish to work through the problem themselves, Weinman provides the following values for comparison:

$$P = .41, \quad S = .53, \quad \hat{\eta} = 1.41*$$

The quantiles and the ordered distances when the S^{-1} matrix is used are shown in Table 10. These should be plotted as before.

When we compare the results obtained when the Identity and the S^{-1} matrices are used, we find that D is the largest effect in both and that SK and SDP are large in both. But DNP and SDNK which are not large when the Identity matrix is used (i.e., when "distance" is defined as the sum of the performance measures squared) are large when the S^{-1} matrix (i.e., the sum of the reciprocal of the variances of each performance measure with their covariation taken into account) is used. Weinman (1979) explains this as follows: DNP is "found" by the S^{-1} matrix because the value of Y_2 at DNP is much larger than that of Y_1. A similar statement is true at SDNK, where the sign of Y_1 is negative while the sign of Y_2 is positive."**

OTHER COMPOUNDING MATRICES

Roy, Gnanadesikan, Srivastava (1971, p 103) state, regarding the choice of the compounding matrix, A: "A truly multivariate situation cannot usually be fully described by a single one-dimensional representation and,

---

*It is happenstance that the values of $\hat{\eta}$ are the same for the Identity and S^{-1} matrices in this example.

**The results from this analysis cannot be related to anything in the real world since the data for Y_2 were contrived.

TABLE 10.   DATA REQUIRED TO PLOT THE ORDERED
DISTANCES FOR THE S^{-1} MATRIX

Ranks	Quantiles	Ordered Distances	Sources
1	.06	0.10	SDK
2	.14	0.12	SDN
3	.21	0.17	DP
4	.27	0.22	SPK
5	.33	0.34	DNPK
6	.39	0.35	N
7	.46	0.46	SDPK
8	.52	0.50	SN
9	.58	0.67	SNP
10	.64	0.75	SPK
11	.71	0.80	PK
12	.78	0.82	SD
13	.85	0.83	NPK
14	.92	0.89	DN
15	1.00	0.96	DNK
16	1.08	0.97	DK
17	1.16	1.16	SDNP
18	1.25	1.23	SNK
19	1.34	1.27	SP
20	1.45	1.35	SNPK
21	1.55	1.39	SDNPK
22	1.68	1.74	P
23	1.81	1.82	NP
24	1.95	2.18	NK
25	2.12	2.19	K
26	2.32	2.34	S
27	2.55	3.45	SDP
28	2.83	3.94	SDNK
29	3.20	4.14	SK
30	3.75	6.68	DNP
31	4.82	8.36	D

therefore, it should be recognized that, for any given prob-
lem, it is always advisable to try several different com-
pounding matrices in the calculation of the squared
distances $\{d_i\}$. In fact, it would often be desirable not
only to try out several measures of size or squared dis-
tances which are positive semi-definite quadratic forms,
but also to try different kinds of measures of size or
distance functions. The problem of the choice of A is
under continuing study and a preliminary report of certain
findings of the study may be found in Wilk et al (1962."
They point out that any interpretation of the results must
be made conditional upon the choice of the A matrix.

SECTION VIII

THE PLACE FOR REPLICATION IN ECONOMICAL
MULTIFACTOR RESEARCH

To psychologists, replicating an experimental design is as natural as breathing and occurs just as unconsciously. Unfortunately, it is a costly procedure when economy in multifactor research is paramount; fortunately, it is not always necessary. It is important that the investigator recognize the different situations in which replication may or may not be desirable or required.

REPLICATING REQUIRED LESS IN THE EARLY rHASES OF RESEARCH

Before an unbiased model of a response surface has been derived, replicating a design is generally not cost effective. During the screening phase, it is more productive to use any extra data-collection effort to add new points to the fractional factorial than it is to repeat original conditions. For example, if one were to replicate the conditions of a Resolution III design, the only additional information that would be obtained would be an estimate of the error variance. If instead of replicating, however, data were collected to complete a second properly selected Resolution III design, although no error would be externally estimated, main and two-factor interaction effects could be isolated, a considerable increase in information. Replicating to obtain an error estimate is not justified in the screening phase since other techniques can be initially employed to get a rough estimate of essentially the same information. For example, an approximation of error might be obtained from the left over sources of variance when main effects do not saturate the design; when they do, order statistics can provide an internal error estimate. If trivial effects are present, they can provide a source of "discovered" replication. Precision in multifactor screening design is derived from "hidden" replication. These terms and methods have been discussed in other reports by Simon (1973, 1979).

REPLICATION USEFUL TO ESTABLISH PSYCHOLOGICAL CONFIDENCE

Even when an investigator conscientiously tries to include all factors in his experiment, he may not be successful because of cost and time pressures, or because he has not yet identified the source of variance. Critical subject characteristics, which in the holistic approach to behavioral research is just as important as equipment or environmental factors, are often difficult to identify. While an investigator may believe that he has considered all the critical subject factors, he may wish to test this assumption by running several subjects on the same conditions. While this is often referred to as "replication," it is so only to the extent that subject characteristics have indeed included all major sources of subject variance.

The prudent investigator, however, will test this assumption. He will repeat all or parts of the experiment using two or more individuals that are presumably "identical," being, in fact, identical on the potentially critical factors already identified. By taking this precaution, the investigator will compare -- not average -- the results from several subjects to see if the same critical effects are revealed in essentially the same order of magnitude among individuals. If two or more presumably identical subjects perform differently, it warns the investigator of the possibility that other unidentified factors are operating. Simon (1977a) discusses various patterns of responses that might occur and possible explanations for them. If the results prove to be essentially identical, then and only then should they be averaged.

PARTIAL REPLICATION FOR ERROR ESTIMATES

When the response surface is being derived, it becomes important to estimate how well the regression model fits the real data. The lack of fit variance is compared in an F-test with some external estimate of error variability. Box and Hunter (1958) propose that it would be economical and often sufficient to obtain this estimate by only replicating at the center of the central-composite design.

Replication at the center of a central-composite design may not provide a powerful enough test of significance (Simon, 1976a, pp 16-18) in all cases. Furthermore, if it is feared that variability may increase away from the center of the design, the estimate at the center may be too small to correctly assess the coefficients of the second-order polynomial. Since we do not really know whether the variance is or is not homogeneous over the total surface, we may wish to replicate at points other than the center. It is not necessary, however, to replicate the entire design. Partial replication can increase the precision of our estimated effects, add more degrees of freedom to the error estimate, and improve the evaluation of the regression model.

A number of people have proposed partially duplicated designs. Daniel (1978) discusses designs proposed by Clatsworthy (1973) for the partial replication of two-way layouts. Patel (1963) describes the partial duplication of two-level fractional factorial designs. Dykstra (1960) proposes several plans for reproducing certain experimental conditions of the central-composite, second-order response surface designs. As a general principle with central-composite designs, replicating either the star or the cube portion (or a fraction thereof) is sufficient for partial replication. Replicating the star increases the precision of estimates away from the center of the experimental space.

Combining Data from Partial Replication of a 2^{k-p} Design

When a sub-fraction of a Resolution V fractional factorial is used for partial replication, Box (1966) provides a simple method of combining the old and new data. The Resolution V (or V-) design indicates that with the original block of data, all critical main and two-factor interaction effects have been isolated from one another. In the sub-fraction, however, this is not the case.

New estimates of each effect can be made by using the following equation to combine the old and new data:

$$\text{New Estimate} = \begin{pmatrix} \text{Old estimate} \\ \text{from first} \\ \text{block} \end{pmatrix} + \frac{\alpha\, n_2}{n_1 + n_2 p} \cdot \begin{bmatrix} \text{Effect} & \text{Effects} \\ \text{of} & \text{of iso-} \\ \text{String} & \text{lated} \\ & \text{sources} \\ & \text{combined} \end{bmatrix}$$

Where α = plus or minus one, which corresponds to the sign of the aliased effect

n_1 = number of observations in first set of data

n_2 = number of observations in second set of data

p = number of aliased effects in string

Let us use this to combine the data from a 2_V^{8-2} design with data from a 2^{8-4} design. Given:

First set: n_1 = 64 conditions

Isolated effects, /A/ = 5, /BC/ = 2, /GH/ = 1

Second set: n_2 = 16 conditions

Effect of string, (A - BC - GH) = 2.5

p = three effects in string

Substituting in the equation we get:

$$\hat{A} = /A/ + \frac{+1 \times 16}{64 + 16 \times 3} [(2.5 - /A/ - /BC/ - /GH/]$$

$$\hat{A} = 5 + \frac{16}{112} [2.5 - (5 - 2 - 1)]$$

$$\hat{A} = 5 + \frac{1}{7} [-0.5] = 5 + (-.0714) = 4.93$$

and

$$\hat{BC} = 2 + \frac{-1 \times 16}{64 + 16 \times 3} [-0.5]$$

$$\hat{BC} = 2 + (-\frac{1}{7}) (-.5) = 2.07$$

and

$$\hat{GH} = 1.07$$

The error variance of the new estimate (Var_N) is:

$$Var_N = \frac{4\sigma^2}{n_1} [\frac{n_1 + n_2 (p - 1)}{n_1 + n_2 p}]$$

which for this example would be:

$$Var_N = \frac{4\sigma^2}{64} [\frac{64 + 16 (3 - 1)}{64 + 16 \times 3}]$$

$$Var_N = \frac{\sigma^2}{16} [\frac{96}{112}] = \frac{6}{7} \sigma^2$$

Thus the additional observations reduced the error variance of the new estimate to approximately 86 percent of the original variance.

REPLICATION TO ESTABLISH CONFIDENCE LIMITS

At the end of a research program, after presumably two or three equipment configurations have been selected based on the results of the multifactor experiments, the investigator may wish to make a more careful comparison of these final choices, or to evaluate them against some earlier version to decide whether or not the replacement is worthwhile. To make

a more precise estimate of the differences in means, particularly if there is an interest in drawing the conclusion of "no (practical) difference," the investigator will wish to replicate these few experimental conditions.*

In judging whether the differences are important or not, the investigator must trade-off the costs of making a wrong judgment (Type I and II errors) against the costs of more data collection. Rather than do the conventional test of statistical significance, the investigator may wish to establish a <u>confidence interval</u>, the limits between which a hypothesis can be considered tenable and outside which, untenable for a certain probability value. It is interesting to note that Cochran and Cox (1957) wrote in their book on experimental designs for hypothesis testing: "On the whole, however, tests of significance are less frequently useful in experimental work than confidence limits" (p 5). Box, Hunter and Hunter (1978) in their book on <u>Statistics for Experimenters</u> say: "Significance testing in general has been a greatly overrated procedure, and in many cases where significant statements have been made it would have been better to provide an interval within which the value of the parameter would be expected to lie" (p 109).

Some equations for estimating confidence limits are shown in Table 11. Note that in the subscript for t, in addition to "d.f." (degrees of freedom), we also select the t for a particular α(probability of making a Type I error). In some textbooks, this α is replaced by $\alpha/2$. For example, in the text by Cochran and Cox (1966), α is used. In the text by Box, Hunter, and Hunter (1978) $\alpha/2$ is used. How can these apparently different equations be reconciled? The difference lies in the t-table one intends to use. In Cochran and Cox, the t-table in the back of the book shows the t-values for a two-tailed test and the probability value used to enter the table is the proportion of <u>both</u> ends of the distribution summed. In Box, Hunter and Hunter, the t-table in the back of the book (p 631) shows the t-values for a one-tailed test and the probability value used to enter

---

*We didn't replicate much or at all when we had a great many conditions to investigate. As we reduce the number of conditions being examined, we begin to replicate more and more.

TABLE 11. EQUATIONS FOR ESTIMATING CONFIDENCE LIMITS

1 − α LIMIT FOR MEAN
(Two-tailed test)

Mean $\pm$ $t_{(n-1, \alpha)} \sigma/\sqrt{n}$

where $\sigma = \sqrt{\Sigma (y - \bar{y})^2/n - 1}$

1 − α LIMIT FOR MEAN DIFFERENCE (Paired)

Mean Diff. $\pm$ $t_{(d.f., \alpha)}$ (S.E. Mean diff.)

where S.E. Mean diff. $= \sqrt{\dfrac{(d - \bar{d})^2}{(n-1)n}} = \dfrac{\sigma}{\sqrt{n}}$

where $d = Y_{a1} - Y_{b1}$

where degrees of freedom are appropriate to t

1 − α LIMIT FOR MEAN DIFFERENCE (Unpaired)

Mean Diff. $\pm$ $t_{(d.f., \alpha)}$ $\sigma\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}$

where $\sigma = \sqrt{\dfrac{(n_A - 1)\sigma_A^2 + (n_B - 1)\sigma_B^2}{(n_A - 1) + (n_B - 1)}}$

the table is the proportions at one end of the symmetrical
distribution. Thus, if an investigator wants the t-value
for 15 degrees of freedom and a .05 probability of making
a Type I error in a two-tailed test, he would look up the
t in the α = .05 column in Cochran and Cox's book and
α/2 = .025 in the Box, Hunter, and Hunter book. In both
cases, t = 2.131. If he wanted a .05 probability of making
a Type I error in a one-tailed test, he would look up t in
the α = .10 column in Cochran and Cox's book and α/2 = .05
in the Box, Hunter, and Hunter book. In both cases, t =
1.753.

Hader and Grandage (1958), in an excellent discussion of
simple and multiple regression analysis, explain and
illustrate how to derive the confidence limits for regression
coefficients and predicted values for a multivariate
situation.

REPLICATION TO ESTABLISH PERFORMANCE LIMITS

Traditional confidence limits are intended to provide a
basis for pinpointing the value of a parameter, a mean or a
mean difference. There is also a need to replicate a single
condition at the very end of an experimental program to
answer the very practical question: Given this device,
between what limits will the group of people, theoretically from
the same population, likely to perform? We are concerned with
the one and ninety-ninth percentile levels here rather than
the fiftieth. While the results from the entire research
program (along with the holistic approach) should make
prediction of these limits quite accurate, there is still
the need to determine them empirically.

SECTION IX

THE SIGNIFICANCE OF TESTS
OF STATISTICAL SIGNIFICANCE

In recent years, the test of statistical significance has become the most widely used analysis performed by behavioral scientists doing controlled experiments. The results from these tests have often become the primary criteria used by experimenters, teachers, and editors alike in evaluating the importance of an experimental study. In fact, however, tests of statistical significance as ordinarily applied by psychologists provide very little useful information in general and often lead to erroneous conclusions in specific cases.

Baken (1966) wrote: "The test of statistical significance in psychological research may be taken as an instance of a kind of essential mindlessness in the conduct of research" (p 436). Lykken (1968) wrote: "Statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published" (p 151). Coats (1970) wrote: "Most graduate schools of education still require students to take what may be one of the most irrelevant learning experience in their entire educational career. The requirement is the study of inferential statistics" (p 6). Cronbach (1975) argued that "the time has arrived to exorcise the null hypothesis" (p 124) and Shulman (1970) admonished that "the time has arrived for educational researchers to divest themselves of the yoke of statistical hypothesis testing" (p 389). Carver (1978) suggests that "the complete abandonment of statistical significance testing in the training of doctorial students in education research should be seriously considered" (p 396).

The problem naturally is not with the test of statistical significance itself but with the way which too many psychologists misuse and misinterpret the test. Even when used correctly, however, it provides little information of a practical value.

In Table 12 are a list of fallacies and facts regarding tests of statistical significance, particularly as they have been applied in the behavioral sciences. This list has been culled from a number of critical papers by Baken (1966), Carver (1978) Kleiter (1969), and others referenced in these three papers. The reader is urged to savor the original papers.

TABLE 12.   FACTS AND FALLACIES REGARDING
TESTS OF STATISTICAL SIGNIFICANCE

Fallacy: The p value associated with the F-ratios in the
test of statistical significance (TSS) indicates
the probability that the observed effect is due
to chance.

Fact: The p value is the probability that an effect of
the observed size could be obtained if, in fact,
it was a certainty that chance is operating.

●

Fallacy: The p value from the TSS is the probability
that the same result would be obtained if the
experiment were replicated. It is an indica-
tion of the reliability of the result.

Fact: Reliability depends on how well critical factors
are controlled between experiments. It cannot
be predicted by any statistics.

●

Fallacy: The level of statistical significance is
inversely related to the probability that the
research hypothesis is correct (e.g., a .05
significance level, if reached, makes the
chances .95 that the scientific hypothesis is
true).

Fact: The entire study could be invalid no matter what
probability level is obtained in the TSS.
Validity is not a statistical concept, but
depends on the adequacy of sampling and the
quality of the data collection and analysis.

●

Fallacy: Using a TSS enables the scientist to make more
objective decisions regarding the rejection of
the null hypothesis.

Fact: The scientific decisions are still totally sub-
jective. The TSS helps make only statistical
decisions of little practical value. Since an
investigator is expected to use his judgment
to plan and design experiments, why is it so
sinful for him to use it to interpret the data?

●   (Continued)

(Table 12 continued)

Fallacy: The more subjects in the experiment, the more faith one can have that the null hypothesis, significant at the p = .05 level, should be rejected.

Fact: Just the reverse. It takes a larger difference to get the same p-value with ten subjects than with 100.

•

Fallacy: Rejecting a null hypothesis at a reasonable statistical significance level is ordinarily required before one can claim support for a research hypothesis.

Fact: One can almost always get statistical significance by increasing the number of subjects, selecting another p-level, changing from a two-tailed to a one-tailed test and so forth. A failure to obtain a significance level may have nothing to do with the hypothesis and everything to do with sloppy research.

•

Fallacy: A statistically significant factor is important.

Fact: With a large enough N, the difference required for significance may be, for all practical purposes, trivial.

•

Fallacy: A failure to get statistical significance indicates that the factor (effect) is not important.

Fact: One may not have obtained statistical significance because one used too few subjects or because one did a sloppy experiment. Importance has to do with the size of a valid effect.

•

Fallacy: When performance is measured in artificial units, TSS helps to interpret the results.

Fact: TSS is not enough. It is necessary to have some outside anchor point to evaluate the practical significance of the magnitude of the effect.

•

(Table 12 continued)

Fallacy: If a difference is statistically significant, it
          indicates that the members of one group performed
          better than the members of the other group(s).

Fact: The test only compares means, not groups. The
          groups could overlap considerably.

                              •

Fallacy: If the mean of Group A is higher than the mean of
          Group B, then a statistically significant result
          infers that Group A performed better on average
          than Group B.

Fact: This depends on the hypotheses being tested.
          Most psychologists traditionally have used a
          two-tailed, non-directional test. This only
          indicates that the observed difference could have
          occurred with a certain probability if chance had,
          in fact, been operating.

     Psychologists have been prone to misinterpret (and
therefore misuse) tests of statistical significance. Fur-
thermore, they have failed to recognize the extremely limited
information that is derived from the test when it is perfor-
med properly. Too frequently the test is not performed
properly. Too often the hypothesis being tested is not the
one the experimenter really wants, but is the only one with
which he is familiar. The significance of several types of
hypotheses can be tested: loose or sharp, Fisherian or
Neyman-Pearson, one or two-tailed, directional or non-direc-
tional. Psychologists have overwhelmingly limited themselves
to sharp, Fisherian, non-directional, two-tailed tests,
whether the information obtained was what they wanted or not.
To make matters worse, among psychologists there has been
little consistency as to what the "error" term should consist
of. Simon (1976b) after analyzing the experiments published
over 14 years in the journal Human Factors found the error
variance was some composite of 17 different combinations of
classes of factor designations (i.e., equipment, subject, and
temporal sources, main and interaction effects). The choice
affects the outcome of the test.

In the light of this confusion surrounding the use and misuse of the test of statistical significance, its role as an "automatic decision maker" for those who wish to escape their responsibility as data interpreters is not justified. In practice, the investigator can perform a number of analyses that will supply the information he needs to decide whether an effect is likely to exist or not (Simon, 1977a; 1977b).

The test can be used most effectively when the investigator is interested in the <u>absence</u> of a practical difference among conditions (see Section X) rather than in the existence of a difference. When the interest is primarily that of detecting differences, then the weaknesses of the test (enhanced by characteristics of behavioral research) come into play. Cochran and Cox (1957, p 5) point out another "useful property of a test of significance" which "is that it exerts a sobering influence on the type of experimenter who jumps to conclusions on scanty data, and who might otherwise try to make everyone excited about some sensational treatment effect that can well be ascribed to the ordinary variation in his experiment." This circumstance is more likely to occur when only one or two factors are being investigated <u>in vacuo</u> than when they are examined in a larger (holistic) context. In the same vein, with a holistic philosophy, the "ordinary variations" in behavioral experiments are likely to be reduced considerably when investigators actively seek to account for the critical sources of variance rather than trying to hide them within a massive uneconomical replication effort.

Cochran and Cox continue by saying: "On the whole, however, tests of significance are less frequently useful in experimental work than confidence limits." In this regard, Box, Hunter, and Hunter (1978, p 109) state: "Significance testing in general has been a greatly overworked procedure, and in many cases where significance statements have been made it would have been better to provide an interval within which the value of the parameter would be expected to lie."

The test of statistical significance should not be totally ignored. However, it must be put into proper perspective, used cautiously and properly, and given a low priority among a number of other techniques that are available to aid the investigator in his decisions regarding the presence and importance of an effect.

SECTION X

DETERMINING THE PROBABILITY OF ACCEPTING THE
NULL HYPOTHESIS WHEN IN FACT IT IS FALSE
(Applications to the interpretation of screening studies)

Since most psychological experiments are planned to
discover an effect, a difference, or a correlation, an
investigator, applying inferential statistics, is more
interested in rejecting the null hypothesis than in
accepting it. In many cases, in deciding whether the effect
observed in the sample data is real or not, he focuses
almost totally on the risk of making a Type I error and
ignores the risk of making a Type II error. That is, he
tries to minimize the chances of saying there is a
difference when, in fact, there isn't. This orientation has
caused many investigators to ignore the risk of accepting
the null hypothesis (i.e., saying there is no effect, no
difference, no correlation) when, in fact, it is false
(Type II error).

There are circumstances when, as a practical matter,
an investigator should decide to accept or reject a hypo-
thesis by weighing both risks, and more important, there
are circumstances when he should suspend judgment, i.e.,
make no decision until additional evidence has been
collected.*

There are situations, however, in which the option of
suspending judgment cannot be exercised. The investigator
must make a decision on the basis of inadequate evidence.
This can be found in an application of the lack of fit test
when a central-composite design is being employed.

By way of example, let us look at the results from a
study (North and Williges, 1971, Table 5) in which the
investigators made a decision to stop collecting more data
on the basis of their test for lack of fit. Their analysis
of variance table with the lack of fit test is shown in
Table 13. The investigators refused to reject the null
hypothesis because the probability value of the F-ratio
for the lack of fit test was only $p = 0.15$ and they had
established the 0.05 significance level as the critical
cut-off point. Still with only three degrees of freedom in
the term used for "error" (Replications), the test was not
a very powerful one. Clearly this was a situation in which
the evidence made it difficult to make a firm decision

---

* Hays (1963, Chapter 9) has an excellent discussion on
hypothesis testing and interval estimation from the point
of view of the behavioral scientist.

TABLE 13.   REPRODUCTION OF NORTH AND WILLIGES' ANALYSIS
OF VARIANCE FOR NUMBER OF CORRECT LOCATIONS

Source	%	df	Variance	F	P
REGRESSION	(.488)	(4)	.90	5.97	.05
Focus	.106	1	.78	5.18	.05
Density	.331	1	2.44	16.19	.01
Visual Angle	.011	1	.08	0.52	
TV Lines	.040		.30	2.00	
RESIDUAL	(.511)	(25)	.15		
Blocks	.001	2	.00	0.068	
Lack of Fit	.493	20	.18	4.301	.15
Replications	.017	3	.04		
TOTAL	(1.000)	(29)			

regarding whether or not there was a fit.   On the one hand,
a p = 0.15 is not small enough for most psychologists to
reject the null hypothesis (although I am not sure why this
must be as strong a rule as it tends to be).   On the other
hand, there is, in fact, a sizeable lack of fit in this
example, for it accounts for nearly 50 percent of the total
performance variance while the two "significant"
experimental factors account for only 33 percent and 11 per-
cent of the total variance.

This example illustrates the dilemma that can occur
when making a Type I error is the only risk considered.   The
purpose of the study was to develop an equation that best
approximated the performance data.   The lack of fit test was
done to see if this had been accomplished or if more data
needed to be collected.   Not rejecting the null hypotheses
-- whether or not it was verbalized that a final judgment
was being suspended -- resulted in a de facto acceptance of
the null hypothesis since no further data was collected to
resolve any ambiguity. That this was not the best solution
becomes clearer when the risk of making a Type I error
(p = 0.15 in this case) is compared with the risk of making
a Type II error.   Calculations show that the risk of saying
that the fit is adequate when, in fact, it isn't, is
p = 0.69.   Thus by failing to reject the null hypothesis

because they did not want to risk (p = .15) making a Type I error, they had accepted a larger risk (p = .69) of making an error of the second kind, implicitly <u>accepting</u> the null hypothesis by not collecting more data. Had they been aware of the size of the second probability, it is presumed the investigators would have continued to collect more data to improve the fit. As it stands, the derived equation probably yields biased predictions.

WEIGHING THE RISKS IN SCREENING DESIGNS

Calculating the risk of making a Type II error can be particularly useful when screening design are involved. Once the results have been obtained, ordered, and plotted on normal (or half-normal) probability paper (see Simon, 1977a), the investigator must decide where to draw the line between the effects that are probably real and those that are probably due to chance. From the plots and from the data itself, the investigator will ord narily find it easy to make decisions regarding the very large and the very small effects. But there are marginal effects between the two extremes about which decisions are difficult to make. Daniel (1976, p 416) has this to say about these marginal effects:

> "The dropping of factors from further consideration after an early 'screening' experiment is sometimes justified. But it is also a very common source of serious errors. The 2^{8-4} may be of higher power than much of the experimenter's previous work, but some care must still be taken to avoid 'Type II' errors. Such errors can be particularly treacherous with multi-factor experiments. For example, one factor may be clearly dominant over the others taken singly. But it may be that three or four of the dropped factors, if varied together, would have had as big an effect as the one called dominant."

When the investigator is faced with the difficult decision of what to do with the marginal sources of variance, several options are open to him. He may decide that since the effect of each marginal variable is small, if he omits it, any equation based on the more obvious variables will account for the major chunk of the performance variance and that any bias from the omitted variables will be of little practical importance. With limited resources, his main concern initially is to account for the larger portion of the total performance variance and to worry about marginal sources at the time may not be cost-effective. If one thinks of experiments within the paradigm of a multi-factor research program (rather than as an experiment), then the investigator expects to have the opportunity to refine his

equations later and ... the marginal cases when
more resources ... ... multifactor
studies, a framework ... ... the smaller studies
that are performed later and related.

It is difficult ... limit the size of
the investigation, the investigator want to include as
many marginal sources as possible in the next stage of his
research. He must ... to keep or exclude
a source of variance ... the statistical risks of
being wrong, or what ... type I and errors.* The
investigator sets his Type I error when he sets the
statistical level ... the null hypothesis
will be rejected. ... other characteris-
tics of his experiment ... calculate the risk
of making a Type II ...

CALCULATING THE RISK ... RRORS

A Type II error ... acceptance of the null
hypothesis is, in fact ... (1963) describes how
to estimate the probability ... a Type II error when
a t-test of statistical ... employed. There are
also published tables ... estimating these
risks but these are ... situations involving
a small number of degrees ... in the numerator and/or
for probability levels 0.1 and ... . Because one may
wish to estimate the risk ... error under a wider
range of conditions, a method ... calculating these values
is described below.**

Here are ... construct the operating
characteristic ... establish the
risk of making ...

An example based ... Rule 12 is given to
illustrate the procedure ...

...

A. Determine ... m for each
    variance ...

*If the investigator ... as Daniel suggests,
that several factors ... al but marginal
effects and in combination ... portion of variance
on a level with the ... ffects, he may wish to
evaluate their ...

**Dr. Thomas ... Company, provided the
steps required to make this ... ion.

B) Select the significance level, α, that the investigator will accept as the risk of making a Type I error when rejecting the null hypothesis.

Example: A) $df_1 = 20$, $df_2 = 3$; B) α = .05.

Step II.

A) Obtain a table of critical values of the F-distribution.

B) For the appropriate degrees of freedom, find the F-value in the table associated with the probability, β = (1-α).

Example: A) One of the most complete tables of critical values of the F-distribution was published by D. B. Owen, Handbook of Statistical Tables, Wesley, 1952, Section 4.1, pp 63-87. A portion of the table from his page 66 is shown here.

Degrees of Freedom For Numerator

Υ	13	14	15	18	20	24		
.500	2.0773	2.0858	2.0931	2.1104	2.1190	2.1321	.500	
.750	9.4399	9.4685	9.4934	9.5520	9.5813	9.6255	.750	
.900	60.903	61.073	61.220	61.567	61.740	62.002	.900	
.950	244.69	245.37	245.95	247.32	248.01	249.05	.950	1
.975	979.85	982.54	984.87	990.36	993.10	997.25	.975	
.990	6125.9	6142.7	6157.3	6191.6	6208.7	6234.6	.990	
.995	24504	24572	24630	24767	24836	24940	.995	
.500	1.3672	1.3727	1.3771	1.3879	1.3933	1.4014	.500	
.750	3.3997	3.4091	3.4098	3.4208	3.4263	3.4345	.750	
.900	9.4145	9.4230	9.4247	9.4358	9.4413	9.4496	.900	
.950	19.419	19.424	19.429	19.441	19.446	19.454	.950	2
.975	39.421	39.426	39.431	39.442	39.448	39.456	.975	
.990	99.422	99.427	99.432	99.443	99.449	99.458	.990	
.995	199.42	199.43	199.43	199.44	199.45	199.46	.995	
.500	1.2021	1.2071	1.2111	1.2205	1.2252	1.2322	.500	
.750	2.4523	2.4537	2.4552	2.4585	2.4602	2.4626	.750	
.900	5.2097	5.2047	5.2003	5.1898	5.1845	5.1764	.900	
.950	8.7286	8.7148	8.7029	8.6744	8.6602	8.6385	.950	3
.975	14.305	14.277	14.253	14.196	14.167	14.124	.975	
.990	26.983	26.923	26.872	26.751	26.690	26.598	.990	
.995	43.271	43.171	43.065	42.680	42.778	42.622	.995	
.500	1.1305	1.1349	1.1386	1.1473	1.1517	1.1583	.500	
.750	2.0827	2.0828	2.0829	2.0829	2.0828	2.0827	.750	
.900	3.8853	3.8765	3.8689	3.8515	3.8443	3.8310	.900	
.950	5.8910	5.8732	5.8578	5.8219	5.8025	5.7744	.950	4
.975	8.7148	8.6836	8.6565	8.5971	8.5599	8.5109	.975	
.990	14.306	14.248	14.198	14.079	14.020	13.929	.990	
.995	20.602	20.514	20.438	20.287	20.007	20.030	.995	

Degrees of Freedom for Denominator

B) w... .05) = .95, which
is the p value be... r 20 and 3 degrees
of freedom. ... s of freedom for
the numerator d...) An arrow points
to the correct ...

Step III.

A) Prep... ...ed as follows:

I		IV
B Probability		λ

B) ... ...the values

Example: ... ...

$$\frac{\lambda}{1}$$

Step IV.

A) ... ... II by copying
the ... ...s provided in
Owen's ... ... than the first

B) ... ...ing the original
... ... subsequent
...

C) ... ... λ^2.

Example: λ

...			1.00
			1.29
...	...	...	1.88
.500	...	...	2.66

* ...

Step V.

A) ... ... .0 it is
necess... ... at $p = (1 - \beta)$
with ... ... Therefore if
we want ... ... 20 and 3 degrees
of freedo... ... s tables (page
76) the ... ... 750 with 3 and 20
... ...iprocal.

Example: For β = .250, 20 and 3 df; look up .750, 3 & 20 df,
           = 1.4808 and take the reciprocal = .6753.

B)   Complete the four-column table.

Example:

β	F	λ^2	λ
.250	.6753	12.82	3.58
.100	.4201	20.61	4.54

Step VI.

A)   Plot the values on a graph with the probabilities
     (β) on the ordinate and

$$\lambda \ (= \frac{\sigma 1}{\sigma 2} = \text{the } \underline{\text{square root}} \text{ of } F)$$

on the abscissa.

B)   Connect the points in a smooth curve. (A
     certain degree of imprecision here can't be
     avoided unless a great many more values are
     plotted.)

Example: See Figure 6 for a plot of the operating
           characteristic curve to be used with the
           data in   le 13. For the particular
           F-valu   ... risk of making a Type II
           error c 1 be estimated. (Note that the
           squar oot of the F-value is used to
           enter the table.)

           Thus, to estimate the risk of a Type II
           error when the F equals 4.301 (Table 13,
           Lack of Fit), we look up the square root
           of F (2.07) and find the associated
           probability (when α = .05 and there are
           20 and 3 degrees of freedom) to be .69.
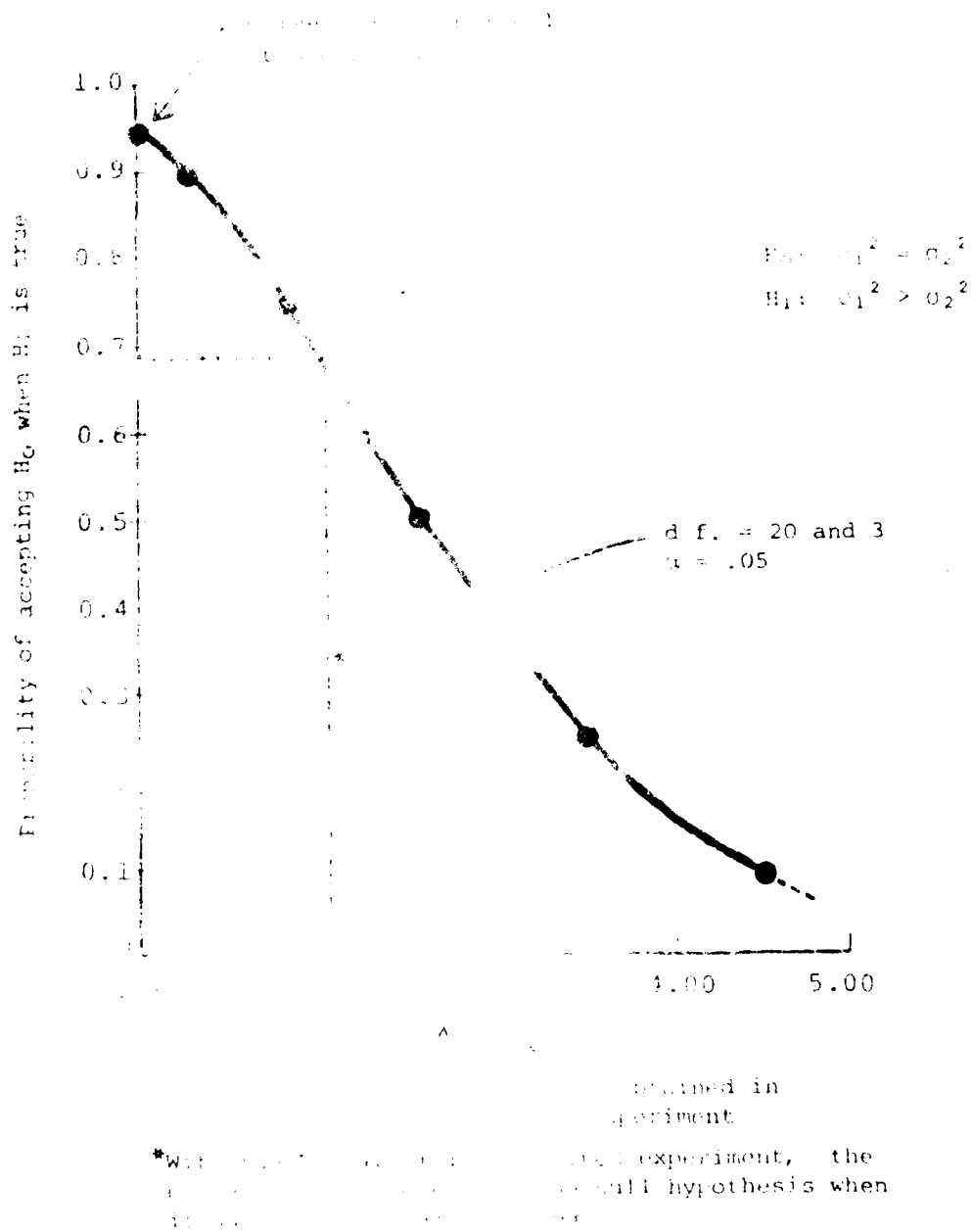
Figure 6. Operating Characteristic Curve for ... in Table 13.

SECTION XI

TESTING NON-ADDITIVITY IN EXPERIMENTAL
DATA FROM A LATIN SQUARE DESIGN

Psychologists have made considerably more use of Latin square designs than is warranted, often disregarding or being unaware of the conditions that must be met before the data can be considered unbiased and tests of statistical significance valid. The popularity of this design rests with its application to what is sometimes referred to as a "within subject" design. A single subject is tested on all experimental conditions presented to them sequentially and by having the same number of subjects as experimental conditions, the order of presentation can be counter-balanced.* With eight subjects and eight conditions presented in counterbalanced orders, the Latin square is an 8 x 8 matrix. This design is actually a fractional factorial in which the effects of three factors -- Conditions, Subjects, and Trials -- are being examined in a two-dimensional space. In an 8 x 8 design, there are 63 degrees of freedom partitioned into 7 for Conditions, 7 for Subjects, 7 for Trials, with 42 left -- actually a confounding of interactions -- to be treated as error. This design is efficient therefore only when there is, in fact, no interaction of any kind among the three sources. In human performance studies, this is often an untenable assumption (without considerable preparation ahead of time) since it will be invalid if the subjects learn at different rates (subject x trial interaction). Then too interactions are often found between subjects and conditions due to a non-linear reaction of subjects of different abilities to tasks of different difficulties. Because of these inherent dangers, an investigator should obtain a rough estimate of whether or not non-additivity is present to avoid misinterpreting his results.

Tukey (1949; 1955) provides a test for non-additivity for matrices of data in which there is a single observation per cell. If the test for non-additivity is statistically significant, it suggests that the linear-by-linear interaction component is confounded with the main effects and "error" variance. The test is particularly sensitive to the non-additivity that occurs when there is a correlation between a subject's average performance and the rate at which his performance changes relative to the change in group performance.

---

*If there is an odd number of experimental conditions, twice as many subjects are required for counterbalancing.

Following Tukey ... ... , the ... ps required to test
for this non-additivi... ... ... employing a Latin
square design are giv...

TUKEY'S TEST OF NON-ADD... ... ... QUARES

In Table 14, the data ... ... quare with one value
per cell is presented in an ... ... n which rows are
Subjects, Columns are Tr... ... ...tions are distributed
in a counterbalanced mann... ... ...ls.* The circled
numbers in the table relate tha... ...ortion of the
table to the steps in the analys... ...scribed below. To
perform the test for non-additiv... ... a Latin square, the
following steps will be...

1. Obtain the grand mean of ... e original data.

2. Obtain the mean of each row (Subject) and find
   how much each deviates from the grand mean.

3. Do the same for each column (Trial).

4. Do the same for each treatment (Letters).

5. Find the predicted value for each cell by
   fitting the additive model: Grand Mean +
   [the sum of the deviati... for the row,
   column, and treatment for he cell] each
   with the correct signs.

6. Subtract the predicted ... ... observed
   value in each cell co... ... the residual in
   each cell. (Arithmetic ... k: These must
   sum to zero in each row ... lumn and treatment).

7. Subtract the grand mean ... m each predicted
   value, square each diffe... ce and build a
   new array (YY) with thes... alues. (Note:
   If the squared values are large and unwieldy,
   they may be divided by a power of 10 that
   reduced them in size to a whole number with
   the original number of significant figures).

8. Do an analysis of variance on the array of values
   obtained in Step 7 and obtain the <u>interaction</u>
   <u>sum of squares</u> for array YY. (Interaction sum
   of squares equals the total sum of squares minus
   the column sum of squares minus the row sum of
   squares minus the treatment sum of squares).

---

*The data was taken from a Latin square in Davies (1967,
p 194). The example was prepared ... Elizabeth Lage Roscoe.

TABLE 14. EXAMPLE OF TEST OF NON-ADDITIVITY IN A LATIN SQUARE

XX		t_1	t_2	t_3	t_4	$\bar{s}_i$	$\bar{s}_{\Delta i}$ ②
S_1	x	A +11	B +11	D -13	C -11	-3	-2.5
	k	12.5	-7.0	-14.25	-3.25 ⑤		
	(x-k)	-1.5	8.0	1.25	-7.75 ⑥		
S_2		B -6	C +33	A +34	D -14	11.75	12.25
		-8.0	29.5	35.75	-10.25		
		2.0	3.5	-1.75	-3.75		
S_3		D -5	A -4	C -22	B +28	-.75	-.25
		-9.25	5.75	-22.5	23		
		4.25	-9.75	0.5	5.0		
S_4		C -45	D +30	B -10	A -15	-10	-9.5
		-40.25	31.75	-10	-21.5		
		-4.75	-1.75	0	6.5		
$\bar{T}_i$		-11.25	15	-2.75	-3	Grand Mean -0.5 ①	
$\bar{T}_{\Delta i}$ ③		-10.75	15.5	-2.25	-2.5		

Conditions	$\bar{C}_i$	$\bar{C}_{\Delta i}$
A	25.75	26.25
B	-20	-19.5
C	1.75	2.25
D	-9.5	-9.0 ④

X Data

$$\Sigma x^2_{INT} =_X \Sigma x^2_{TOT} - \Sigma x^2_S - \Sigma x^2_T - \Sigma x^2_C$$

⑪

$$\Sigma x^2_{INT} =_X 7444 - 986.5 - 1468.5 - 4621.5$$

$$\Sigma x^2_{INT} =_X 367.5 \quad [\text{d.f.} = 15 - 9 = 6 \quad Var_{Int.} = 61.25]$$

(Continued)

93

TABLE 14 (Continued)

YY

	[k - x]² (7)	(x - k) (6)	Prod 6×7 (9)

A	B	D	C
16?	42.25	175.6	7.6
-1.5	8.0	1.25	-7.75
-?3.5	9?8	219.5	-58.9

B	?	A	D
56.2?	?4?	1314.1	95.1
2.?	?.?	-1.75	-3.75
-1?.5	?7.?	-2299.7	-346.6

D	A	C	B
76.6	39.1	484	552.25
4.25	-9.75	0.5	5.0
315.6	-381.?	242	2761.2

C	D	E	A
1580.1	1040.1	90.25	441.
-4.75	-1.75	0	6.5
-7505.5	-1820.2	0	2866.5

(9a) $\left[\left(\Sigma \textcircled{9} \right) \right]^2 = 369{,}785.$

Step 7 Data

$$\Sigma x^2_{INT \cdot Y} = \Sigma x^2_{TOT} - \Sigma x^2_S - \Sigma x^2_T - \Sigma x^2_C$$

$$\Sigma x^2_{INT \cdot Y} = 3769106 - 1117218 - 147222 - 89443$$

$$\Sigma x^2_{INT \cdot Y} = 1610233 \quad d.f. = 1 \quad Var._{Int.} = 1610233$$

$$\Sigma x^2_{NON-ADD} = \frac{369785}{1610233} = .2296 \quad d.f. = 1 \quad Var. = .23$$

Interaction Σx²	6	367.50		
Non-Additivity	1	.23	.23	.003(NS)
Remainder	5	367.27	73.45	

9. Multiply the value in each cell from Step 7 by its corresponding residual value from Step 6. Sum all of these values, and square that answer (9A).

10. Divide the squared value from Step 9A by the value in Step 8 and obtain the sum of squares for non-additivity with one degree of freedom. The non-additivity variance is the same as its sum of squares.

11. Obtain the interaction sum of squares for array XX. (See Step 8 for the equation with which to extimate the interaction sum of squares.) The degrees of freedom for this interaction equals the total degrees of freedom, (N - 1), minus three times the number of degrees of freedom for treatments, (T - 1).

12. Subtract the sum of squares for the interaction of the XX array (Step 11) to obtain the remainder sum of squares. The remainder degrees of freedom equals the degrees of freedom for the interaction XX array sum of squares minus one.

13. Obtain the remainder variance by dividing its sum of squares (Step 12) by its degrees of freedom (Step 12).

14. The F-test for non-additivity is made by dividing the non-additivity variance (Step 10) by the remainder variance (Step 13). This will be evaluated in the conventional manner using a standard F-distribution table for 1 and (N - 3T + 1) degrees of freedom.

If the F-value is statistically significant or the non-additivity sum of squares accounts for a sizable proportion of the total sum of squares, then we must be prepared to reject the hypothesis of additivity and recognize that this assumption (fundamental when using a Latin square) is not being met.

What to Do?

Some may be tempted, in the face of non-additivity, to ignore its one degree of freedom and to use the mean square for the balance as the "error" term for significance tests. Tukey (1949, p 237) does not recommend this. First of all, he points out, it is more practical to express results in additive terms since they tend to apply over a breader region. Second, if the "error" variance is non-normally distributed, the balance variance "unduly inflates the

apparent significance of the other mean squares". In the presence of a large non-additivity term, Tukey suggests an examination of the data to determine whether the non-additivity is due to one or more unusually discrepant values or whether it is due to analysis in the wrong form.

Tukey (1949) proposes a graphic method to help this decision.* For small amounts of data the technique can be ambiguous. The reader is referred to the paper for further enlightenment. If it is decided that the non-additivity was due to analyzing the data in the wrong form, then transformation of the data is in order.

GENERAL FORM OF TUKEY'S NON-ADDITIVITY TEST

Tukey (1955) provides a general procedure for performing the non-additivity test for any design. The steps are simple and are listed below to enable the reader to gain insight into what he is actually doing. To test for non-additivity in any factorial-type design with a single measure per cell:

1. Obtain the residual for each cell (one per cell). $[x]$

2. Square the predicted values for each cell and treat those values as original data, and obtain the residuals for that set of data for each cell. (Nothing will be changed if a constant is added to the predicted values prior to squaring.) $[y]$

3. Obtain the sum of cross-products of values obtained in steps 1 and 2 found in corresponding cells. $[\Sigma xy]$

4. Obtain the sum of squares of values obtained in step 1. $[\Sigma x^2]$

5. Divide the sum obtained in step 3 by the sum obtained in step 4. This is the estimated coefficient (i.e., least squares estimate) for non additivity. $[\Sigma xy/\Sigma x^2]$

6. Square the sum obtained in step 3 and divide by the sum obtained in step 4 to obtain the sum of squares for non-additivity with one degree of freedom (which is partitioned from the residual sum of squares in the ANOVA table and tested as any other source of variance in the ANOVA.) $[(\Sigma xy)^2/\Sigma x^2]$

---

*The method was developed for a row-by-column table and will need to be modified to be applied to the Latin square.

SECTION XII

HOW TO INCLUDE FACTORS WITH MORE THAN TWO LEVELS IN A SCREENING DESIGN

Screening studies are usually 2^{k-P} fractional factorial designs. Two levels are used since the goal is to identify the critical candidate factors and not to obtain a functional relationship. If these two goals are separated, the economy of large-scale multifactor experiments can ordinarily be increased (Simon, 1973, 1977a). Once the critical factors have been found, additional levels may be added to this smaller number, if necessary, to study a more complex function and response surface.

There are circumstances, however, when an investigator may wish to study some factors at more than two levels during the screening phase. For example, if the factor is a qualitative (categorical) one, there is no clear-cut method of selecting only two out a larger number of different categories. While the ordinary procedure would be to select the categories believed likely to represent extremes in performance -- this judgment being based on life experiences or preliminary studies -- this approach may sometimes be considered too tenuous. In that case, the investigator may wish to include several categories in his original screening study to obtain information regarding each one and some clues regarding how they might interact with other factors. If the factor is a quantitative one and is believed to relate to performance with a U-shaped function, selecting two points at which performance differences are likely to be close to maximum may be less accurate than is desired. If there is sufficient uncertainty, the investigator may wish to insure himself by studying three or four levels along the dimension.

Whatever the reason for wanting to include three or four levels, the investigator must weigh the advantages against the loss in data-collection economy and the increased difficulty in interpreting the results. Certainly the practice of having more than two levels should be used sparingly during the screening phase.

METHODS OF INCLUDING MORE THAN TWO LEVELS

There are several ways in which factors with more-than-two-levels might be introduced into a screening design. One way would be to include the three-or-more level factor outside the basic screening plan. This means that the screening design would be repeated at every level of the new factor. This reduces the economy of the experiment, but if the added data-collection effort can be tolerated, it might be used if the factor is an important one and

disordinal interacti ns between it and other factors are
suspected. A second way of introducing the multilevel
factor would be to use $2^{4.6}$ fractional designs that already
exist (Conner and ____, 1961). This approach, however, is
the least attractive since these designs are seldom econom-
ical and not really useful for screening purposes. A third
method can be employed when the basic screening design is
not saturated, that is, when it has the capacity to isolate
more independent effects than the design is required to
measure. In this case, there may be enough space to intro-
duce a three- or four-level factor into the screening design.
This has the advantage of maintaining the economy of the
effort during this early phase of the program. It has the
disadvantage of increasing the problems of data interpre-
tation.

INCLUDING A FOUR-LEVEL ____ ____ SCREENING DESIGN

Addelman (1962, p ____, ____ ____ Cochran and Cox, 1957,
pp 273-274) show how three two-level factors can be
replaced by one four-level factor without confounding main
effects. This is how it works:

Conditions	Original Factors				New Factor Levels
	A	B	AB		
1	-	-	+	———————→	α
2	.	.	.	———————→	β
3	.	+	.	———————→	λ
4	.	.	.	———————→	δ

Note that all three columns (or factors) are orthogonal to one
another, as are the columns in a screening design. However,
while the experimenter may freely select any two columns of
a screening design that he wishes, the third must be the
column representing the interaction of the first two, even
though in the screening design this may be aliased with a
main effect or ____ ____ ____ ____ only three such
columns, only four unique patterns of signs will be found
in the rows, i.e., --+, +--, -+-, and +++. The procedure is
to designate each as one of four levels, α, β, λ, and δ, of
a new four-level factor. Here is how the method would
used to create a 4×2^4 fractional factorial design.

Step 1: Select an experimental design of adequate size.

Since we must use three columns of an orthogonal design
for each four-level factor, a screening design capable of
estimating seven main effects is required for the 4×2^4
fractional factorial in this example. A 2^{7-4}_{III} design will be
used, made up of the following eight experimental conditions:

Conditions   New Factor Labels

	A	B	C	D	E	F	G
1	−	−	−	+	+	+	−
2	+	−	−	−	−	+	+
3	−	+	−	−	+	−	+
4	+	+	−	+	−	−	−
5	−	−	+	+	−	=	+
6	+	−	+	−	+	−	−
7	−	+	+	−	−	+	−
8	+	+	+	+	+	+	+

| | A | B | C | AB | AC | BC | ABC |

Step 2:   Select any two columns (i.e., factors) plus a
          third that includes the generalized interaction
          as its alias.

     We will use factor (columns) A and B. This forces us
to use factor D since D is aliased with the AB interaction
in the Resolution III design. In a Resolution IV design,
the third effect would be the isolated string containing
the appropriate interaction.

Step 3:   Replace each unique sign pattern (in a row of signs
          of the selected factors) with a symbol representing
          each different level of a four-level factor.

A	B	D		X
−	−	+	⟶	α
+	−	−	⟶	β
−	+	−	⟶	λ
+	+	+	⟶	δ
−	−	+	⟶	α
+	−	−	⟶	β
−	+	−	⟶	λ
+	+	+	⟶	δ

The levels of factor A are designated categories α, β,
λ, δ for a qualitative factor or (for example)-3, -1, +1,
+3, respectively, if factor A is a quantitative factor.

Step 4: Write the complete 4 x 2^4 main effects design.

	A		E	F	G
1)					-
2)				+	+
3)				-	+
4)				-	-
5)				-	
6)		+	+		-
7)				+	-
8)		+		-	+

All five main effects are orthogonal to one another.*
It is a Resolution III design. With this "main-effect"
(i.e., Resolution III) design, the assumption that all
higher-order effects are negligible must be valid or else
the results will be biased. If that assumption cannot be
made, additional data must be collected according to the
strategy to be used in all screening studies.

_____

*Principle of Proportional Frequencies (Plackett, 1946).
This principle states that "...a necessary and sufficient
condition that the main effects estimates of two factors be
uncorrelated is that the levels of one factor occur with
each of the levels of the other factor with proportional
frequencies." Furthermore, it also states that "...for
main effects to be uncorrelated with two factor interaction
effects each combination of the levels of two factors must
occur with the levels of another main effect with propor-
tional frequencies." The same proportional relationships
hold for fractional factorial, 2^{k-p} designs, characteris-
tically used in screening studies.

COLLAPSING FOUR-LEVEL FACTORS TO THREE-LEVEL FACTORS

Three-level factors are not created directly from two-level factors. Instead, a four-level factor must first be created through replacement and collapsed to become a three-level factor. Addelman (1963, p 61) illustrates this thus:

Two-level factor	Replacement	Four-level factor	Collapsing	Three-Level factor
- - +	$\longrightarrow$	α	$\longrightarrow$	α
+ - -	$\longrightarrow$	β	$\longrightarrow$	β
- + -	$\longrightarrow$	λ	$\longrightarrow$	λ
+ + +	$\longrightarrow$	δ	$\longrightarrow$	β

Category δ is changed to Category β. For quantitative factors, the values -1, 0 and +1 could be substituted for the categories α, β, and λ. When the three-level factor is introduced this way into a 2^{p-q} screening design, the Principle of Proportional Frequency guarantees that orthogonality will be maintained.

Let us use the 4×2^4 main effect plan developed in Steps 1 through 4 of the previous section to create a 3×2^4 main effect plan. Collapsing the four-level factor to three levels in the manner shown above would create the following design:

Four-level factor	Collapsing	Three-level factor	Remainder of Two-Level factors			
α	$\longrightarrow$	α	-	+	+	-
β	$\longrightarrow$	β	-	-	+	+
δ	$\longrightarrow$	δ	-	+	-	+
β	$\longrightarrow$	β	-	-	-	-
α	$\longrightarrow$	α	+	-	-	+
β	$\longrightarrow$	β	+	+	-	-
λ	$\longrightarrow$	λ	+	-	+	-
δ	$\longrightarrow$	β	+	+	+	+
		Y	C	E	F	G

Factors

AUGMENTING THE 3 x 2^{k-p} DESIGN

When multilevel factors are developed from 2^{k-p} designs in the manner described, the plans remain "main effect" designs. This means that the assumption -- it must be a valid one -- is made that no interaction effects exist. This assumption is generally not tenable in the behavioral sciences and for this reason we usually insist that main effects be isolated at least from two-factor interaction effects.

Daniel (1976, p 229) describes how this would be done with a 3 x 2^{k-p} design. He proposes that the usual reversal of signs be made with the two-level factors -- as with the "foldover" design -- and that for the three-level factor, the levels would be interchanged in the following manner:

$$\alpha \longrightarrow \beta$$
$$\beta \longrightarrow \alpha$$

$$\lambda \longrightarrow \beta$$
$$\beta \longrightarrow \lambda$$
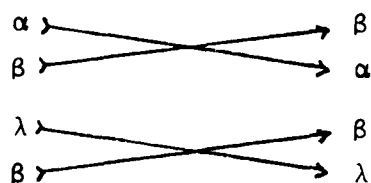
Daniel discusses the analysis of these designs if the multilevel factor is a quantitative variable. Two dummy variables for linear and quadratic effects would be substituted for the three-level factor. With these new "variables", the aliases can be determined (p 226-229).

Because of ambiguity in interpretation, until more experience is obtained, the reader should use these designs cautiously for whatever information or clues that might be obtained without full dependency upon the results.

SECTION XIII

ANALYZING EXTRA-PERIOD CHANGE-OVER DESIGNS

To conserve subjects and to make more precise comparisons among treatments, a common procedure in human engineering experiments is to test the same subject sequentially on a series of different experimental conditions. In practice, however, the intended advantages of this type of design may be outweighed by biases introduced by effects artificially created by the sequential presentation. Many psychologists attempt to overcome this problem by using counterbalanced Latin square designs in which each treatment appears once in every column (period) and once in every row (subjects). When each treatment is arranged so that it precedes and follows every other treatment just once, the design is referred to as a carry-over or change-over design. In addition to being able to isolate the effects of treatments, subjects, and periods, the change-over design enables a residual effect carried over from the treatment given on the previous trial to be isolated from the direct treatment effect. Simon (1974) describes the construction of a number of different change-over designs in his summary of techniques for handling various sequence effects.

In the basic Latin square change-over design, estimates of direct and residual treatments effects are not completely independent, with the precision of residual estimates being lower than the direct. In this case, such designs are more useful when the main reason for isolating residual effects is to provide an unbiased estimate of the direct effects. By adding an additional period and repeating the last treatment of the series given to each subject, a more balanced design can be formed. With this "extra-period design," direct and residual effects are independent of one another and have approximately equal precision.

Examples of how to analyze the basic Latin square change-over design can be found in the statistical literature rather easily (Cochran and Cox, 1957; Federer, 1964). This is not the case for the extra-period change-over design. Lucas (1957) describes the analysis of the extra-period design in the Journal of Dairy Sciences. Patterson and Lucas (1962) provide a description in a Technical Bulletin published by the North Carolina Agricultural Experiment Station. Cochran and Cox (1957) provide the necessary equations but gives no numerical example for this analysis. Because these references may be difficult for a reader to obtain, the steps in the analysis of the extra-period change-over design are described here, along with a numerical example.

ANALYZING THE RESULTS FROM AN EXTRA-PERIOD CHANGE-OVER DESIGN

In Table 15-A, an extra-period change-over design is shown for four conditions (A through D). The balanced 4x4 Latin square arrangement has been supplemented with an additional period (I through V) to orthogonalize direct treatment and residual treatment (a through d) effects. Fictitious performance scores for each of the 20 conditions are given in parentheses.

Some analyses have been completed and the results are given in the margins of the design in Table 15-A:

P_i = sum of all scores made in each period
    e.g., $P_{III} = 8 + 3 + 4 + 2 = 17$

S_i = sum of all scores made by each subject
    e.g., $S_2 = 1 + 6 + 3 + 4 + 2 = 16$

ΣY = sum of all N scores = $\Sigma S_i = \Sigma P_i$

ΣY^2 = sum of all scores after squaring each one

N = total number of conditions = $n_S \times n_P$

n = number of treatments = number of subjects

In Table 15-B, results from additional preliminary analyses are given:

T_i = sum of all scores made for each treatment
    (direct), e.g., $T_D = 5 + 3 + 8 + 4 + 2 = 22$

R_i = sum of all scores made on the period immediately following each treatment i
    (residual), e.g., $R_a = 6 + 4 + 4 + 5 = 19$

$S_i - P_{Ii}$ = sum of all scores for each subject except for that occuring in the first period,
    e.g., $S_2 = 6 + 3 + 4 + 2 = 15$

In addition, we must obtain:

R_i^2 = sum of the residual values after each are squared

$(\Sigma Y - \Sigma P_I)$ = sum of all scores except those made during the first period = $\Sigma(S_i - P_{Ii})$

In this example, since t is even, there is a single Latin square. If t were odd, then a minimum of two squares would be needed to create the desired balance. The letter q equals the number of Latin squares; in our example, q = 1.

TABLE 15. EXTRA PERIOD CHANGE-OVER DESIGN, DATA, AND PRELIMINARY ANALYSES

Part A

SUBJECTS (S)

		1	2	3	4	P_i
	I	$A_{(3)}$	$B_{(1)}$	$D_{(5)}$	$C_{(4)}$	13
	II	$a^B{}_{(6)}$	$b^C{}_{(6)}$	$d^A{}_{(3)}$	$c^D{}_{(3)}$	18
Periods (P)	III	$b^D{}_{(8)}$	$c^A{}_{(3)}$	$a^C{}_{(4)}$	$d^B{}_{(2)}$	17
	IV	$d^C{}_{(7)}$	$a^D{}_{(4)}$	$c^B{}_{(2)}$	$b^A{}_{(4)}$	17
	V	$c^C{}_{(2)}$	$d^D{}_{(2)}$	$b^B{}_{(1)}$	$a^A{}_{(5)}$	10
S_i		26	16	15	18	$\Sigma Y=$ 75

Residual(R)
Treatments(T)
Performance(Y)

$\Sigma Y^2 =$ 353

$N =$ 20

Part B

T_i : 18 12 23 22

		a	b	c	d
R_i	:	19	19	10	14

$\Sigma R_i^2 = 1018$

		1	2	3	4
$S_i - P_{I_i}$	:	23	15	10	14

$(\Sigma Y - \Sigma P_I) = 62$

N = Total number of observations
n = t = number of conditions (treatment)
q = Number of Latin squares = 1

In Table 16, the analysis of variance is completed. Sums of squares and degrees of freedom and variances are obtained for the following sources of variance: total, periods, subjects, direct and residual treatment effects, and error. This differs from a conventional analysis only in the calculations for the direct and residual treatment effects, where direct effects must be isolated from its overlap with subjects and the residual effects must be isolated from an overlap with periods. These "overlaps," or correlations, can be detected in the design shown in Table 15A, e.g., subject 1 has two treatment C's but only one of all other conditions, subject 2 has two treatment D's but only one of the others, and so forth, while no residual appears in period 1. The magnitude of the direct and residual treatment effects are those obtained after eliminating the correlated portion.

In Table 17, additional equations (found in Lucas, 1957) are supplied for the analysis when t is odd and two Latin squares (q = 2) are involved. The only new term here is:

$$Q_i = \text{sum of all values in each Latin square}$$

When there are two, then the sum of squares, degrees of freedom and variance must be estimated for Latin squares. New sum of squares for periods, subjects, and error are calculated first by the equations in Table 16, and then corrected by subtracting the sum of squares for squares from each other source, as shown in the equations in Table 17. These new or modified values are then included in an analysis of variance along with the values of the direct and residual treatment effects which are calculated using the same equations as shown in Table 16. The symbol $\Sigma x^2_{P/Q}$ indicates that it is the sum of squares for period within Latin squares; any others that include the /Q indicate this same change in source of variance.

LIMITATION IN THE USE OF EXTRA-PERIOD CHANGE-OVER DESIGNS

These designs assume a linear model, as indicated in the tables. Therefore, they would not be adequate if it is suspected that there is an interaction between direct and residual treatment effects. The linear model is applicable only if the residual effect for each treatment remains essentially constant regardless of the treatment that follows. The DxR interaction implies that the magnitude of a residual effect will vary depending on which treatment follows which treatment. In many human performance studies, particularly those involving motor tasks, one cannot assume that the linear model is valid.

Simon (1979a) had suggested that change-over designs might provide a new and economical way of studying transfer of training. The few designs in the literature that are based on a DxR model are uneconomical, requiring too many observations to properly balance only a relatively few treatments. Efforts to develop more economical designs for this interaction model were not successful (Simon, 1979b, Section III).

**TABLE 16.   ANALYSIS OF VARIANCE OF EXTRA-PERIOD
CHANGE-OVER DESIGN WHEN t IS EVEN AND
q IS ONE LATIN SQUARE**

			$\Sigma x^2 i$	d.f.	Var.
TOTAL:	$\Sigma Y^2 - \dfrac{(\Sigma Y)^2}{N} = \Sigma x_T^2$			[N-1]	
	$353 - \dfrac{(75)^2}{20} = 353 - 281.25 =$	Σx_T^2	71.75	19	3.78
PERIODS:	$\dfrac{\Sigma Pi^2}{n} - \dfrac{(\Sigma Y)^2}{N} = \Sigma x_P^2$			q[n]	
	$\dfrac{1171}{4} - 281.25 =$	Σx_P^2	11.5	4	2.88
SUBJECTS:	$\dfrac{\Sigma Si^2}{n+1} - \dfrac{(\Sigma Y)^2}{N} = \Sigma x_S^2$			q[n-1]	
	$\dfrac{1481}{5} = 281.25 =$	Σx_S^2	14.95	3	4.98
DIRECT:	$\Sigma [(n+1)(\Sigma T_i) - \Sigma S_{t_i} - \Sigma Y]^2 / qn(n+1)(n+2) = \Sigma x_D^2$			[n-1]	
Ta:	$[(5)(18) - 18 - 75]^2 = [-3]^2 = \quad 9$				
Tb:	$[(5)(12) - 15 - 75]^2 = [-30]^2 = \quad 900$				
Tc:	$[(5)(23) - 26 - 75]^2 = [14]^2 = \quad 196$				
Td:	$[(5)(22) - 16 - 75]^2 = [19]^2 = \dfrac{361}{}$				
	$\Sigma = \dfrac{1466}{4 \cdot 5 \cdot 6} = \Sigma x_D^2$		12.22	3	4.07
RESIDUAL:	$\dfrac{\Sigma R_i^2}{qn} - \dfrac{(\Sigma Y - \Sigma P_I)^2}{qn^2} = \Sigma x_R^2$			[n-1]	
	$\dfrac{1018}{4} - \dfrac{(62)}{4^2} = 254.5 - 240.2 =$	Σx_R^2	14.25	3	4.75
ERROR:	$\Sigma x_T^2 - \Sigma x_P^2 - \Sigma x_S^2 - \Sigma x_D^2 - \Sigma x_R^2 = \Sigma x_E^2$			{(qn-2)(n-1)}	
	$71.75 - 11.5 - 14.95 - 12.22 - 14.25 =$	Σx_E^2	18.83	6	3.14
MODEL:	$\Sigma x_T^2 = \Sigma x_D^2 + \Sigma x_R^2 + \Sigma x_P^2 + \Sigma x_S^2 + \Sigma x_E^2$				

TABLE 17.   ADDITIONAL EQUATIONS FOR ANALYSIS OF VARIANCE
OF EXTRA-PERIOD CHANGE-OVER DESIGN WHEN t IS
ODD AND q IS TWO LATIN SQUARES

ADD TO ANALYSES IN TABLE 16

LATIN SQUARES (Q)   d.f.

$$\Sigma x_Q^2 = \Sigma Q_i^2 - \frac{(\Sigma Y)^2}{N}$$ [q-1]

MODIFY ANALYSES IN TABLE 16

PERIODS WITHIN SQUARES (substitute for PERIODS)

$$\Sigma x_{P/Q}^2 = \Sigma x_P^2 - \Sigma x_Q^2$$ [qn]

SUBJECTS WITHIN SQUARES (substitute for SUBJECTS)

$$\Sigma x_{S/Q}^2 = \Sigma x_S^2 - \Sigma x_Q^2$$ [q(n-1)]

ERROR WITHIN MULTIPLE SQUARES (substitute for ERROR)

$$\Sigma x_{E/Q}^2 = \Sigma x_E^2 - \Sigma x_Q^2$$ [(qn-2)(n-1)]

MODEL $$\Sigma x_T^2 = \Sigma x_D^2 + \Sigma x_R^2 + \Sigma x_Q^2 + \Sigma x_{P/Q}^2 + \Sigma x_{S/Q}^2 + \Sigma x_{E/Q}^2$$
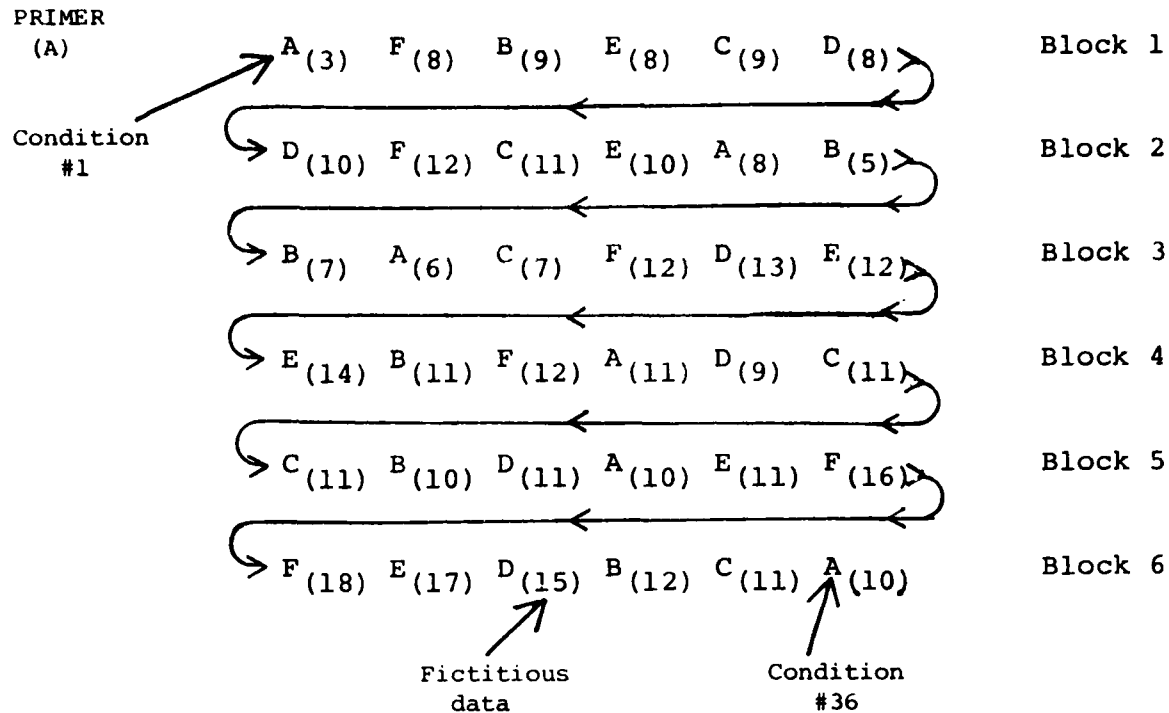
SECTION XIV

ANALYZING SERIALLY-BALANCED SEQUENCE DESIGNS

A serially-balanced sequence design is a type of change-over design used to isolate direct treatment effects from residual effects carried over from a preceding treatment. It differs from the change-over designs in the preceding section since it is balanced over a series of replicated conditions run by a single subject rather than among a group of subjects with subjects, trials, and treatments arranged in a Latin square format. Sampford(1957) describes how these S.B.S. designs are constructed and analyzed. Simon (1974), in a report summarizing techniques for handling sequence effects, describes the construction of these designs, but not the analysis. Since Sampford's explanation of the analysis may be difficult for some psychologists to follow, the method of analyzing serially balanced sequence designs (Type 1, k = 1) is given here along with a numerical example. The Type 1 design, like the extra-period change-over design, balances the treatment so that direct and residual treatment effects are orthogonal. However, residual effects remain confounded with blocks. Isolating block effects from estimates of residual effects creates the only complication in the analysis. The k = 1 indicates that a single sequence is used.

METHODS OF ANALYSIS

An example of a serially balanced sequence design is shown in Table 18 along with fictitious performance data. Although the conditions are arranged in a Latin square, they are actually to be presented to a single subject serially beginning with the condition in the upper left-hand corner, moving across the row, back to the left end of the next row, and so forth until the lower right-hand condition is reached. The first condition in parentheses is called a "primer" and is not used in the calculations. It is there to provide a residual effect for the next condition. Thus, performance on the second A is the result of the combined effects of the direct treatment effect of A in that period, plus any effect carried over from the primer A in the preceding period, plus any block effect. Similarly, the performance level for the next period in the sequence is the result of the direct effect of condition F, the residual effect from the preceding condition A, and the effect of Block 1.

The symbols and equations required to perform the analysis are shown below and should be calculated in the order given.

TABLE 18. SERIALLY BALANCED SERIAL DESIGN
WITH FICTITIOUS DATA*

(Type 1, t=6, k=1)

PRIMER
(A)

$A_{(3)}$ $F_{(8)}$ $B_{(9)}$ $E_{(8)}$ $C_{(9)}$ $D_{(8)}$ Block 1

Condition
#1

$D_{(10)}$ $F_{(12)}$ $C_{(11)}$ $E_{(10)}$ $A_{(8)}$ $B_{(5)}$ Block 2

$B_{(7)}$ $A_{(6)}$ $C_{(7)}$ $F_{(12)}$ $D_{(13)}$ $E_{(12)}$ Block 3

$E_{(14)}$ $B_{(11)}$ $F_{(12)}$ $A_{(11)}$ $D_{(9)}$ $C_{(11)}$ Block 4

$C_{(11)}$ $B_{(10)}$ $D_{(11)}$ $A_{(10)}$ $E_{(11)}$ $F_{(16)}$ Block 5

$F_{(18)}$ $E_{(17)}$ $D_{(15)}$ $B_{(12)}$ $C_{(11)}$ $A_{(10)}$ Block 6

Fictitious
data

Condition
#36

---

*These fictitious data were created by weighting the Treatments, Direct
and Residual, and Blocks in this way.

Let Direct Treatment A = 1 Add 1 if it follows A and Add 1 if in Block 1
                      B = 2   "   2   "   B   "   2   "   Block 2
                      C = 3   "   3   "   C   "   3   "   Block 3
                      D = 4   "   4   "   D   "   4   "   Block 4
                      E = 5   "   5   "   E   "   5   "   Block 5
                      F = 6   "   6   "   F   "   6   "   Block 6

Thus, Condition 1 = [A=1]+[A=1]+[Block 1=1] = 3

      Condition 16 = [F=6]+[C=3]+[Block 3=3] = 12

      etc.

No error component was used.

1. N = total number of observations, $y_i = t^2$, where t is the number of treatments.

2. G = sum of all scores = $y_1 + y_2 + \ldots y_n$

3. M = grand mean = G/N

4. T_i = sum of all scores for each treatment, e.g., $T_{A1} + T_{A2} + \ldots T_{At} = T_A$

5. B_i = sum of all scores for each block, e.g., $B_{11} + B_{12} + \ldots B_{1t} = B_1$ (ignoring possible residual effects that might be present).

6. R_i = sum of all scores for each residual, e.g., $R_{A1} + R_{A2} + \ldots R_{At} = R_A$ (ignoring possible block effects that might be present).

> The scores for the residuals of Factor i are those assigned to the treatments that <u>follow</u> the Factor i treatment. For example, in the <u>second</u> block, the score containing the residual for treatment E will be (8), and for treatment A will be (5). Note that the residual of the primer treatment is included here, e.g., 3 for primer treatment A, as well as 8 for first treatment A.

7. Σx_G^2 = total sum of squares = $\Sigma Y^2 - \dfrac{(G)^2}{N}$

8. Σx_T^2 = treatment sum of squares = $\dfrac{\Sigma T_i^2}{t} - \dfrac{(G)^2}{N}$

9. Σx_B^2 = block sum of squares ignoring possible residual effects. = $\dfrac{\Sigma B_i^2}{t} - \dfrac{(G)^2}{N}$

The only complicated part of the analysis is where residual effects are adjusted by removing any overlap with block effects. The following explains the steps for doing this.

10. R_i' = residual total adjusted for block effects =

$$tR_i - G + B_{i-1} - B_i$$

> Important note: The numerical order for residuals does NOT correspond to their alphabetical order. Instead, each residual (and its treatment) receives the number of the block in which it appears first. Thus, if i = 3 in our example, the R_3 that goes with B_3 will be R_B since both the residual and the direct effect of Treatment B is the first one in Block 3. Where i for R_i is 2, R_2 will be R_D.

11. The following is a "cookbook" description of the steps required to estimate the residual effects isolated from block effects (R^n). See the numerical example, page 105.

Constructing the coefficient matrix

a. Develop a coefficient matrix with t columns and t rows where t equals the number of treatments. Identify the columns as follows:

> Write the letter a above the first column. Then write subsequent letters in alphabetical order, first forward and then reversed, enough to cover the remaining columns in a symmetrical pattern. For example, when t = 6, the letters for the columns would be a, then b, c, d, c, b; when t = 7, the letters would be a, then b, c, d, d, c, b.

b. Write the numerical value equal to (t^2-2) along the main diagonal of the matrix (upper left to lower right).

c. Write the number 1 in each row to the right and left of the diagonal. Where there is no space to the left (first row) or right (last row), the number 1 is placed at the opposite end of that row. Put zeros in the remaining cells.

$$
\begin{array}{cccccc}
a & b & c & d & c & b \\
\end{array}
$$

$$
\begin{bmatrix}
(t^2-2) & 1 & 0 & 0\ldots & 0 & 1 \\
1 & (t^2-2) & 1 & 0\ldots & 0 & 0 \\
0 & 1 & (t^2-2) & 1\ldots & 0 & 0 \\
\vdots & & \vdots & & \vdots & \\
0 & 0 & 0 & 0\ldots & (t^2-2) & 1 \\
1 & 0 & 0 & 0\ldots & 1 & (t^2-2)
\end{bmatrix}
$$

Writing, solving, and inverting the normal equations

d. Write the normal equations as the sum of the products between each row coefficient and its corresponding columnar term (letter). Only [t/2 + 1], rounded down, of the t equations will be unique, the exact number being the number of different letters (terms) formed in Step 11a.

e. Set the equation in the first row equal to one. Set all of the remaining rows equal to zero.

f. Solve for unknown terms using the usual algebraic processes. Start with the bottom equation and work up, substituting the value of each term as it is determined. Be careful to maintain the correct arithmetic signs.

g. Use the values obtained for each unknown to write the first row of the inverted matrix by placing each value under the appropriate term in the column designated in Step 11a.* Complete the remaining rows of the inverted matrix by horizontally rotating the first row to the right, one column at a time for each succeeding row.

h. To solve for the residual effects with block effects removed, R", we multiply the vector of R_i' values by the inverse coefficient matrix. To do this, the elements of the first row of the coefficient matrix are multiplied by the corresponding elements of the column of R_i' values and summed. This sum is divided by the common denominator, the result will be the first element of the R_1'' column, i.e., $R_1'' = A_{11}R_1' + A_{12}R_2' + \ldots + A_{1n}R_n'$. Similarly, multiplication of elements from the second row of the coefficient matrix with the corresponding elements of R_i' values will give the second element in the R_i'' vector (R_2''). To simplify this, it helps to write the t R_i'' values in order above the columns of the coefficient matrix. Then R_i'' will be the sum of the cross product of the R_i' value and the corresponding coefficient in each row, where the value of i in R_i'' corresponds to row i in the matrix.

12. B_i'' = block effects with residual effects removed =

$(tB_i - G - tR_i'' + tR_{i+1}'')/t^2$ (where $R_{t+1}'' = R_1''$)

13. $(\Sigma x_B^2 + \Sigma x_{R''}^2)$ = composite sum of squares for block and for residual eliminating blocks = $\Sigma R_i R_i'' + \Sigma B_i B_i''$

14. $\Sigma x_{R''}^2$ = residual sum of squares eliminating blocks =

$(\Sigma x_B^2 + \Sigma x_{R''}^2) - \Sigma x_B^2$ = (Step 13) - (Step 9)

---

*Sampford (1957, p 302) provides solutions for the first row of inverse matrices for t = 6 to 10 inclusive. These are given in Table 19.

TABLE 19.   FIRST ROW VALUES OF INVERSE MATRIX

t =	6	7	8	9	10
	19601	105937	7380481	39424240	4517251249
	-577	-2255	-119071	-499121	-46099201
	17	48	1921	6319	470449
	-1	-1	-31	-80	-4801
	17	-1	1	1	49
	-577	48	-31	1	-1
	[665280]*	-2255	1921	-80	49
		[4974529]*	-119071	6319	-4801
			[457351680]*	-499121	470449
				[3113516718]*	-46099201
					[442598424000]*

[*Denominators]   (From Sampford, 1957, p 302)

15. Σx_E^2 = composite total sum of squares less treatment sum of squares, residual sum of squares (eliminating blocks), and block sum of squares (ignoring residuals) =

$$\Sigma x_G^2 - \Sigma x_T^2 - \Sigma x_{R''}^2 - \Sigma x_B^2 =$$

(Step 7) - [(Step 8)+(Step 13)]

16.   Sources, sums of squares, and degrees of freedom.

Source	Σx^2	d.f.
Total	Σx_G^2	t^2-1
Treatments: Direct	Σx_T^2	$t-1$
Residual	$\Sigma x_{R''}^2$	$t-1$
Blocks	Σx_B^2	$t-1$
Error	Σx_E^2	t^2-3t+2

NUMERICAL EXAMPLE*

This numerical example follows the steps set forth in the "Method of Analysis" section. The full design for six treatments and the data on which this analysis is based are given in Table 18.

1. $N = t^2$ $t = 6$, $N = 36$

2. $G = 3+8+9+8+9+10+12...+11+10 = 378$

3. $M = 378/36 = 10.5$

4. $T_A = 3+8+6+11+10+10 = 48$

   $T_B = 9+5+7+11+10+12 = 54$

   $T_C = 60$, $T_D = 66$, $T_E = 72$, $T_F = 78$

5. $B_1 = 3+8+9+8+9+8 = 45$

   $B_2 = 10+12+11+10+8+5 = 56$

   $B_3 = 57$, $B_4 = 68$, $B_5 = 69$, $B_6 = 83$

6. $R_A = 3+8+5+7+9+11 = 43$

   $R_B = 8+7+6+12+11+11 = 55$

   $R_C = 61$, $R_D = 67$, $R_E = 73$, $R_F = 79$

---

7. $\Sigma x_G^2 = [3^2+8^2+9^2+....15^2+12^2+11^2+10^2] - \dfrac{(378)^2}{36} = 345$

8. $\Sigma x_T^2 = \dfrac{(48)^2+(54)^2+(60)^2+(66)^2+(72)^2}{6} - \dfrac{(378)^2}{} = 105$

9. $\Sigma x_B^2 = \dfrac{(45)^2+(56)^2+(57)^2+(68)^2+(69)^2+(83)^2}{6} - \dfrac{(378)^2}{36} = 145$

---

10. $R_A = R_1$, $R_B = R_3$, $R_C = R_5$, $R_D = R_2$, $R_E = R_4$, $R_F = R_6$

   $B_{1-1} = B_0 = B_6$, $B_{2-1} = B_1$, etc.

   $R_1' = R_A' = 6(43) - 378 + 83 - 45 = -82$

   $R_2' = R_D' = 6(67) - 378 + 45 - 56 = +13$

   $R_3' = R_B' = -49$ $R_4' = R_E' = +49$ $R_5' = R_C' = -13$ $R_6' = R_F' = +82$

---

*This example was prepared by Dr. Howard B. Lee.

11a,b,c.

	a	b	c	d	c	b
	34	1	0	0	0	1
	1	34	1	0	0	0
	0	1	34	1	0	0
	0	0	1	34	1	0
	0	0	0	1	34	1
	1	0	0	0	1	34

$\}$ (Repeats)

11d.

$$34a + 2b = 1$$
$$a + 34b + c = 0$$
$$b + 34c + d = 0$$
$$2c + 34d = 0$$

11e.

11f.

1    $2c = -34d;\ c = -17d$

2    $b = -34(-17d) - d = +577d$

3    $a = -34(+577d) - (17d) = -19601d$

4    $34(-19601d) + 2(+577d) = -665280d = +1$

$$d = \frac{-1}{665280}.$$ Substitute the numerator for d in the other equations. Since the denominator is constant to all terms, it will be held out until later.

11g.

	−82	+13	−49	+49	−13	+87	← R_i'
	a	b	c	d	c	b	

$$\frac{1}{665280}
\begin{bmatrix}
19601 & -577 & 17 & -1 & 17 & -577 \\
-577 & 19601 & -577 & 17 & -1 & 17 \\
17 & -577 & 19601 & -577 & 17 & -1 \\
-1 & 17 & -577 & 19601 & -577 & 17 \\
17 & -1 & 17 & -577 & 19601 & -577 \\
-577 & 17 & -1 & 17 & -577 & 19601
\end{bmatrix}
\bullet
\begin{bmatrix}
-82 & a \\
+13 & b \\
-49 & c \\
+49 & d \\
-13 & c \\
+82 & b
\end{bmatrix}
=
\begin{bmatrix}
R_1'' \\
R_2'' \\
R_3'' \\
R_4'' \\
R_5'' \\
R_6''
\end{bmatrix}$$

11h. $R''_{1(A)} = \dfrac{(19601)(-82)+(-577)(13)+(17)(-49)+(-1)(+49)+(17)(-13)+(-577)(+82)}{665280} = -2.5$

$R''_{2(D)} = \dfrac{(-577)(-82)+(19601)(13)+(-577)(-49)+(17)(49)+(-1)(-13)+(17)(82)}{665280} = .5$

$R''_{3(B)} = -1.5 \quad R''_{4(E)} = 1.5 \quad R''_{5(C)} = -.5 \quad R''_{6(F)} = 2.5$

12. $B''_1 = [6(45) - 378 - 6(-2.5) + 6(.5)]/36 = -2.5$

$B''_2 = [6(56) - 378 - 6(.5) + 6(-1.5)]/36 = -1.5$

$B''_3 = -.5 \quad B''_4 = .5 \quad B''_5 = 1.5 \quad B''_6 = 2.5$

13. $[\Sigma x^2_{B''} + \Sigma x^2_{R''}] = [(43)(-2.5) + (55)(-1.5) + (61)(-.5) +$
$(67)(.5) + (73)(1.5) + (79)(2.5)] +$
$[(45)(-2.5) + (56)(-1.5) + (57)(-.5) +$
$(68)(.5) + (69)(1.5) + (83)(2.5)] = 240$

14. $\Sigma x^2_{R''} = 240 - 145 = 95$

15. $\Sigma x^2_E = 345 - [105 + 240] = 0$ [Note: While this is so for this fictitious example, the error would ordinarily not be equal to zero.]

16. ANOVA Table

Source	Σx^2	Degrees of Freedom	Variance	F*
Σx^2_T	105	5	21	
$\Sigma x^2_{R''}$	95	5	18	
$\Sigma x^2_{B''}$	145	5	29	
Σx^2_E	0	20	0	

*No test was possible in this fictitious example since there was no error variance.

SECTION XV

DESIGN ECONOMY WHEN EXPERIMENTAL FACTORS
SELECTIVELY AFFECT BI-VARIATE CRITERIA

There can be times when some experimental factors will be
expected to affect one set of criterion measures and other
experimental factors will be expected to affect a different
set of criterion measures. In multifactor, multiple criteria
experiments, if such pairings do occur between independent
and dependent variables with infrequent overlaps, Daniel (1960)
has shown how additional economy can be achieved when 2^k or
2^{k-p} data collection patterns are employed.

Let us suppose that in the AWAVS investigation of par-
ameters for a pilot-training simulation of carrier landings,
the investigator has good reasons to believe that the Flols
display will have a significant effect on vertical deviations
from the glideslope but not the horizontal, that wind gusts
across the flight path will significantly affect horizontal
deviations but not the vertical, and that lag between responses
of the visual and motion systems will affect both measures. In
an investigation to quantify these effects, the pattern of
critical effects from the different sources of variance in a
2^3 factorial design would look like this:

	Source	Criteria Vertical (y_1)	Horizontal (y_2)
FLOLS	A	X	
LAG	B	X	X
	AB	X	
WIND	C		X
	AC*		
	BC		X
	ABC*		

An X is placed in the criterion column affected by the source to
the left. The sources with asterisks affect neither criterion.

It is apparent that the same information would be obtained
were two 2^2 experiments run, one to study the effects of A and B
on the vertical measure, the other to study the effects of B and
C on the horizontal measure. In each case, two main effects
and one two-factor interaction could be estimated correctly.
If this were done, however, no economy would have been achieved
by running the two four-condition experiments rather than a
single eight-condition one.

If we arrange the sources and effects into groups of influence in this manner:

y_1	y_2	-
A	BC	
B	B	AC
AB	C	
		ABC

those familiar with aliasing in fractional factorials will note that the effects are combined as they would be in a 2^{3-1} fractional factorial design:

$$(A + BC)$$
$$(B + AC)$$
$$(AB + C)$$

each aliased pair being associated with one degree of freedom. In this fraction, the defining generator is:

$$I = ABC$$

As such, effect ABC cannot be estimated. However, since neither it nor effect AB is believed to have a critical effect on y1 or y2, no information will be lost. Furthermore, since effects C, AC, and BC are negligible on criterion y1, the aliasing will not bias the estimates of A, B, and AB. Similarly, since A, AB, and AC are believed to have negligible effects on y2, estimates of B, C, and BC will not be biased by being aliased with them, insofar as the second criterion is concerned. By aliasing those effects that do not affect the same performance criterion, we are able to cut the size of the experiment in half. Only the following four (out of eight) experimental conditions are needed to complete this half-replicate design:

$$a, \quad b, \quad c, \quad abc$$

The other half-fraction, I = -ABC, might have been used. This would involve the experimental conditions: (1), ab, ac, bc.

Daniel (1960, pp 266-267) supplies the defining generators for a number of eight- and sixteen-condition designs involving from four to eight factors with different "influence patterns," i.e., 1-2-1, 2-1-2, 2-2-2, 3-0-3, 3-1-3, and 4-0-4. An influence pattern is a notation Daniel uses to describe the independent-dependent factor pairings when only two criteria are involved. The influence pattern for the example described in this section would be: 1-1-1, corresponding to the letters A-B-C respectively. The first term indicates the number of factors that affect only y1; the last term, the number affecting only y2; and the middle term, the number affecting both criteria.

Since one cannot always be certain that the particular influences will occur as assumed, designs should be selected so that the incorrectness of the assumptions might be detected. This would seem to set the requirement that at least Resolution IV designs should be used so that no main effects will be confounded with any other main effects nor with two-factor interactions, which would remain in strings. If the design is not saturated, the characteristics of the three-factor interaction strings may provide clues to the correctness of the assumptions and may also serve as an estimate of error to the extent that they are negligible.

It may seem at this point that we have made a complete circle, beginning with a new technique for effecting economy but ending up with the same size design that one would have used anyway in a conventional screening design with one or more criteria. This is not quite the case. Instead, the technique provides a useful and different way of looking at a problem in experimental design and can be most effective and economical when the influence pattern is well defined. It provides an additional basis for deciding how to assign factors to the experimental design structure, which ones to alias and which to isolate. Furthermore, with clues from initial blocks of data available to verify initial assumptions of influence patterns, the size of the effort required to isolate two-factor interactions in strings for purposes of screening or developing response surfaces will be reduced since certain combinations will not be expected to influence either criterion.

Daniel (1960) proposes a pre-experiment analysis of the variables -- an operation that is already a part of Phase I of the "new paradigm" (Simon, 1977b) when economical multifactor designs are used with single criterion -- "to summarize the experimenter's knowledge and feelings about the effects of each of K factors on each of r Responses." He writes (p. 268):

> A K x R "influence matrix" has been useful
> both in aiding the statistician to under-
> stand the limitations and advantages of the
> experimenter's technical background, and to
> record the experimenter's state of belief
> before the new round of experiments is
> started. Entries of -1, 0, +1 can be used
> to indicate the experimenter's opinions
> about the sign and magnitude of real effects.
> An i can be used to indicate ignorance.

Daniel suggests that it might be helpful if criteria could be classified according to types, for example: 1) those that measure similar properties, in the same units, e.g., vertical deviations from glideslope before and after training; 2) those that measure similar or related properties, not of the same dimensions, e.g., vertical and horizontal deviations from the flightpath; 3) those that are based on qualitatively different properties, e.g., cost and performance on particular simulation configuration.

REFERENCES

Addelman, S. Techniques for constructing fractional replicate plans, J. Amer. Stat. Assoc., 1963, 58, 45-71.

Anscombe, F. J., and J. W. Tukey. The examination and analysis of residuals, Technometrics, 1963, 5, 141-160.

Barnett, V. The ordering of multivariate data, J. Roy. Stat. Soc., Series A, 1976, 139, 318-354.

Bakan, D. The test of significance in psychological research, Psychological Bulletin, 1966, 66, 423-437.

Box, G. E. P. An introduction to response surface methodology, Madison: University of Wisconsin, Department of Statistics, Tech. Rept. #33, 1964.

Box, G. E. P., and D. R. Cox. An analysis of transformations, J. Roy. Stat. Soc., Series B, 1964, 26, 211-252.

Box, G. E. P., and J. S. Hunter. The 2^{k-p} fractional factorial designs, Technometrics, 1961, 3, 311-351; 449-458.

Box, G. E. P., W. G. Hunter, and J. S. Hunter. Statistics for experimenters, N. Y.: Wiley, 1978.

Carver, R. N. The case against statistical significance testing, Harvard Educational Review, 1978, 48, 378-399.

Clatworthy, W. H. Tables of two associate class partially balanced designs, National Bureau of Standards, Applied Mathematics Series, No. 63, 1973.

Coats, W. A. A case against the normal use of inferential statistical models in educational research, Educational Researcher, June 1970, pp 6-7.

Cochran, W. G., and G. M. Cox. Experimental designs, New York: Wiley, 1957 (2nd edition).

Conner, W. S., and S. Young. Fractional factorial designs for experiments with factors at two and three levels, U. S. Govt. Printing Office: National Bureau of Standards Applied Mathematics Series No. 58, 1961.

Cotter, S. C. A screening design for factorial experiments with interactions, Biometrika, 1979, 66, 317-320.

Cronbach, L. J. Beyond the two disciplines of scientific psychology, American Psychologist, 1975, 30, 116-127.

Daniel, C. Use of half-normal plots in interpreting factorial two-level experiments, Technometrics, 1959, 1, 311-341.

Daniel, C. Parallel fractional replicates, Technometrics, 1960, 2, 263-268.

Daniel, C. Applications of statistics to industrial experimentation, N. Y.: Wiley, 1976.

Daniel, C., and F. S. Wood. Fitting equations to data, N. Y.: Wiley-Interscience, 1971.

Davies, O. L. Design and analysis of industrial experiments, (2nd ed.), New York: Hafner, 1967.

DeGray, R. J. Design for interactions, Technometrics, 1968, 10, 389-391.

Draper, N. R., and A. M. Herzberg. On lack of fit, Technometrics, 1971, 13, 231-241.

Draper, N. R., and W. G. Hunter. Transformations: Some examples revisited, Technometrics, 1969, 11, 23-40.

Dykstra, Jr., O. Partial replication of response surface designs, Technometrics, 1960, 2, 185-195.

Dykstra, Jr., O. The orthogonalization of undesigned experiments, Technometrics, 1966, 8, 279-290.

Dykstra, Jr., O. The augmentation of experimental data to maximize /X'X/, Technometrics, 1971, 13, 682-688.

Draper, N. R., and H. Smith. Applied regression analysis, N. Y.: Wiley, 1968.

Federer, W. T. Experimental design: Theory and application, N. Y.: Biometrics, 1964, 20, 168-181.

Gere, J. M., and W. Weaver, Jr. Matrix algebra for engineers, N. Y.: D. Van Nostrand, 1965.

Gnanadesikan, R. Some remarks on multivariate statistical methods for analysis of experimental data, Industrial Quality Control, 1963, 19, 22-6 and 31-2.

Hader, R. J., and A. H. E. Grandage. Simple and multiple regression analyses. In Chew, V. (Ed.) Experimental Designs in Industry, N. Y.: Wiley, 1958.

Hays, W. L. Statistics, N. Y.: Holt, Rinehart, and
    Winston, 1963.

Hebble, T. L., and T. J. Mitchell. "Repairing" response
    surface designs, Technometrics, 1972, 14, 767-779.

Hill, W. J. Statistical techniques for model building.
    Ph.D. Thesis, University of Wisconsin, 1966.

Kleiter, G. The crisis of significance tests in psychology.
    Jahrbuch für Psychologie, Psychotherapie und
    Medizinische Anthropologie, 1969, 17, 144-163.
    (Translated by D. P. Barrett, Royal Aircraft Establish-
    ment Library Translation No. 1649, The R.A.E. Library,
    Q.4 BUILDING, R.A.E. Farnborough Hants, England, June 1972).

Lucas, H. L. Extra-period Latin-square change-over designs,
    Journal of Dairy Sciences, 1957, 40, 225-239.

Lykken, D. T. Statistical significance in psychological
    research, Psychological Bulletin, 1968, 70, 151-159.

North, R. A., and R. C. Williges. Video cartographic image
    methodology, Savoy, Ill.: University of Illinois,
    Aviation Research Laboratory Technical Report
    ARL-71-22/AFOSR-71-8, October 1971.

Owen, D. B. Handbook of statistical tables, N. Y.:
    Additon-Wesley, 1962.

Patel, M. S. Partially duplicated fractional factorial
    designs, Technometrics, 1963, 5(1), 71-83.

Patterson, H. D., and H. L. Lucas. Change-over designs,
    Technical Bulletin No. 147, North Carolina Agricultural
    Experiment Station and United States Department of
    Agriculture, September 1962.

Patterson, H. D. Quenouille's changeover designs,
    Biometrika, 1973, 60, 33-45.

Plackett, R. L. Some generalizations in the multifactorial
    design, Biometrika, 1946, 33, 328-332.

Roy, S., R. Gnanadesikan, and J. Srivastiva. Analysis and
    design of certain quantitative multiresponse
    experiments, New York: Pergamon Press, 1971.

Sampford, M. R. Methods of construction and analysis of
    serially balanced sequences, J. Roy. Stat. Soc.,
    Series B, 1957, 19, 286-304.

Shulman, L. S. Reconstruction of educational research,
    Review of Educational Research, 1970, 40, 371-393.

Simon, C. W. Reducing irrelevant variance through the use
    of blocked experimental designs. Culver City, CA:
    Hughes Aircraft Co., Tech. Rep. No. AFOSR-70-5,
    November 1970a, 65pp (AD776-041).

Simon, C. W. The use of central-composite designs in human
    factors engineering experiments, Culver City, CA:
    Hughes Aircraft Co., Tech. Rep. No. AFOSR-70-6,
    December 1970b, 52pp (AD748-277).

Simon, C. W. Considerations for the proper design and
    interpretation of human factors engineering experiments,
    Culver City, CA: Hughes Aircraft Co., Tech Rep.
    No. P73-325, December 1971, 135 pp.

Simon, C. W. Economical multifactor designs for human
    factors engineering experiments, Culver City, CA:
    Hughes Aircraft Co., Tech. Rep. No. P73-326A, June 1973,
    171 pp. (AD 767-739).

Simon, C. W. Methods for handling sequence effects in human
    factors engineering experiments, Culver City, CA:
    Hughes Aircraft Co., Tech. Rep. No. P74-451A,
    December 1974, 197 pp (AD A006-240).

Simon, C. W. Methods for improving information from
    "undesigned" human factors experiments, Culver City,
    CA: Hughes Aircraft Co., Tech. Rep. No. P75-287,
    July 1975a, 82 pp (AD A018 455).

Simon, C. W. Response surface methodology revisited;
    a commentary on research strategy, Westlake Village,
    CA: Canyon Research Group, Inc., Tech. Rep. No.
    CWS-01-76, July 1976a, 60 pp. (AD A043-242).

Simon, C. W. Analysis of human factors engineering
    experiments: characteristics, results and
    applications, Westlake Village, CA: Canyon Research
    Group, Inc., Tech. Rep. No. CWS-02-76, August 1976b,
    104 pp. (AD A038-184)

Simon, C. W. Design, analysis, and interpretation of
    screening designs for human factors engineering
    research, Westlake Village, CA: Canyon Research Group,
    Inc., Tech. Rep. No. CWS-03-77, September 1977a,
    220 pp. (AD A036-985)

Simon, C. W. New research paradigm for applied experimental
    psychology: a system approach, Westlake Village, CA:
    Canyon Research Group, Inc., Tech. Rep. No. CWS-04-77,
    October 1977b, 123 pp. (AD A056-984)

Simon, C. W. Applications of advanced experimental method-
      ologies to AWAVS training research, Orlando: Naval
      Training Equipment Center Tech. Rep. NAVTRAEQUIPCEN
      77-C-0065-1, CWS-01-78, January 1979. (AD A064-332)

Tukey, J. W. One degree of freedom for non-additivity,
      Biometrics, 1949, 5, 232-242.

Tukey, J. W. Answer to query No. 113, Biometrics, 1955,
      11, 111-113.

Weinman, David G. Personal communication, Hollins College,
      VA, 1979.

Williams, E. J. Experimental designs balanced for the
      estimation of residual effects of treatments,
      Australian Journal of Scientific Research, 1949, 2,
      149-168.

Wilk, M. B., and R. Gnanadesikan. Graphical methods for
      internal comparisons in multiresponse experiments,
      Ann. Math. Statist., 1964, 35, 613-631.

Wilk, M. B., R. Gnanadesikan, and M. J. Huyett. Probability
      plots for the gamma distribution, Technometrics, 1962,
      4, 1-20.

Wood, F. S. The use of individual effects and residuals in
      fitting equations to data, Technometrics, 1973, 15,
      677-695.

DISTRIBUTION LIST

Naval Training Equipment Center Orlando, Florida 32813	60	The Van Evera Library 1 Human Resources Research Organization 300 North Washington Street Alexandria, Virginia 22314
Commander HQ, TRADOC Attn: ATTNG-PA Ft. Monroe, Virginia 23651	3	Library 1 Division of Public Documents Government Printing Office Washington, D.C. 20402
Center Library Naval Personnel Research and Development Center San Diego, California 92152	3	Director 2 Training Analysis & Evaluation Group Department of the Navy Orlando, Florida 32813
Dr. Ralph R. Canter U.S. Army Research Institute Field Unit P. O. Box 16117 Fort Harrison, Indiana 46216	1	HumRRO/Western Division/Carmel Office 1 27857 Berwick Drive Carmel, California 93923
Document Processing Division Defense Documentation Center Cameron Station Alexandria, Virginia 22314	12	U.S. Coast Guard (G-P-1/62) 1 400 Seventh Street, SW Washington, D.C. 20590
PERI-OU U.S. Army Research Institute for the Behavioral & Social Sciences 5001 Eisenhower Avenue Alexandria, Virginia 22333	1	JSAS Manuscript Office 3 1200 Seventeenth Street, NW Washington, D.C. 20036
OASD (MRA & L)/Training Room 3B922, Pentagon Washington, D.C. 20301	2	Personnel & Training Research Programs 3 Office of Naval Research (Code 458) Psychological Sciences Div. 800 N. Quincy Street Arlington, Virginia 22217
Dr. Ralph Dusek U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue Alexandria, Virginia 22333	2	National Aviation Facilities 1 Experimental Center Library Atlantic City, New Jersey 08405

DISTRIBUTION LIST (cont'd)

American Psychological Assoc. Psyc. INFO Document Control Unit 1200 Seventeenth Street, NW Washington, D.C. 20036	1	AFOSR/NL (Dr. A. R. Fregley) Bolling AFB Washington, D.C. 20332	1
AFHRL Technology Office Attn: MAJ Duncan L. Dieterly NASA-Ames Research Center MS 239-2 Moffett Field, California 94035	1	Human Factors Society Attn: Bulletin Editor P. O. Box 1369 Santa Monica, California 90406	2
Center for Naval Analyses Attn: Dr. R. F. Lockman 2000 N. Beauregard Street Alexandria, Virginia 22311	1	National Defense University Research Directorate Ft. McNair, D.C. 20319	1
Dr. J. Huddleston Head of Personnel Psychology Army Personnel Research Establishment c/o RAE, Farnborough Hants, ENGLAND	1	Commanding Officer Air Force Office of Scientific Research Technical Library Washington, D.C. 20301	1
OUSDR&E (R&AT) (E&LS) CDR Paul R. Chatelier Room 3D129, The Pentagon Washington, D.C. 20301	1	Dr. D. G. Pearce Behavioral Sciences Division Defense and Civil Institute of Environmental Medicine P. O. Box 2000 Downsview, Ontario M3M, CANADA	1
Dr. Jesse Orlansky Science and Technology Division Institute for Defense Analyses 400 Army-Navy Drive Arlington, Virginia 22202	1	Technical Library OUSDR&E Room 3D122 Washington, D.C. 20301	1
Chief of Naval Operations OP-987H Attn: Dr. R. G. Smith Washington, D.C. 20350	1	Commander Naval Air Systems Command AIR 340F Attn: CDR C. Hutc ns Washington, D.C. 20361	2
Scientific Technical Information Office NASA Washington, D.C. 20546	1	Chief ARI Field Unit P. O. Box 476 Ft. Rucker, Alabama 36362	1

DISTRIBUTION LIST (cont'd)

Chief of Naval Operations OP-115 Attn: M. K. Malehorn Washington, D.C. 20350	1	Dr. Martin Tolcott Office of Naval Research 800 N. Quincy Street Department of the Navy Arlington, Virginia 22217	1
Technical Library Naval Training Equipment Center Orlando, Florida 32813	1	Commander Naval Air Development Center Attn: Technical Library Warminster, Pennsylvania 18974	1
Chief of Naval Operations OP-596 Washington, D.C. 20350	1	Naval Research Laboratory Attn: Library Washington, D.C. 20375	1
Commander Naval Air Test Center CT 176 Patuxent River, Maryland 20670	1	Chief of Naval Education and Training Liaison Office AFHRL/OTLN Williams AFB, Arizona 85224	6
Office of Deputy Chief of Naval Operations Manpower, Personnel and Training (OP-01) Washington, D.C. 20350	1	Dr. Donald W. Connolly Research Psychologist Federal Aviation Administration FAA NAFEC ANA-230 Bldg. 3 Atlantic City, New Jersey 08405	1
Assistant Secretary of the Navy Research, Engineering & Systems Washington, D.C. 20350	1	Chief of Naval Material MAT 08D2 CP5, Room 678 Attn: Arnold I. Rubinstein Washington, D.C. 20360	1
HQ Marine Corps Code APC Attn: LTC J. W. Biermas Washington, D.C. 20380	1	Commanding Officer Naval Education Training Program and Development Center Attn: Technical Library Pensacola, Florida 32509	1
Chief of Naval Operations OP-593B Washington, D.C. 20350	1	Commander Naval Air Systems Command Technical Library AIR-950D Washington, D.C. 20361	1
Scientific Advisor Headquarters U.S. Marine Corps Washington, D.C. 20380	1	Chief of Naval Education and Training Code 01A Pensacola, Florida 32509	1

DISTRIBUTION LIST (cont'd)

Commander Pacific Missile Test Center Point Mugu, California 93042	1	Dr. David C. Nagel LM-239-3 NASA Ames Research Center Moffett Field, California 94035
Commander Naval Air Systems Command AIR 4135B Attn: LCDR J. H. Ashburn Washington, D.C. 20361	1	Federal Aviation Administration Technical Library Bureau Research & Development Washington, D.C. 20590
Commanding Officer Naval Aerospace Medical Research Laboratory Code L5 Department of Psychology Pensacola, Florida 32512	1	Commander Naval Weapons Center Human Factors Branch (Code 3194) Attn: Mr. Ronald A. Erickson China Lake, California 93555
Dr. Thomas Longridge AFHRL/OTR Williams AFB, Arizona 85224	1	CDR Robert S. Kennedy Officer in Charge Naval Aerospace Medical Research Laboratory Box 29407 New Orleans, Louisiana 70189
Dr. Kenneth Boff ARAMRL/HEA Wright Patterson AFB Ohio 45433	1	Dr. J. D. Fletcher Defense Adv. Research Projects Agency (CTO) 1400 Wilson Boulevard Arlington, Virginia 22209
CAPT James Goodson Code L-53 Naval Aerospace Medical Research Laboratory Pensacola, Florida 32512	1	Mr. Robert Wright Aeromechanics Lab (USAAVRADCOM) Ames Research Ctr, MS 239-2 Moffett Field, California 94035
Major Jack Thorpe AFOSR/NL Bolling AFB, D.C. 20332	1	Lt Col Jefferson Koonce USAFA/DFBL USAF Academy, Colorado 80840
Dr. Will Bickley USARI Field Unit P. O. Box 476 Fort Rucker, Alabama 36362	1	CDR Norman E. Lane Code 602 Human Factors Engineering Division Naval Air Development Center Warminster, Pennsylvania 18974
AFHRL/TSZ Brooks AFB, Texas 78235	2	